Language Resources for Public Security Applications

http://www.lrps.amu.edu.pl/

Workshop Programme

14:00 – 14:20 – Opening and introductory presentation by Zygmunt Vetulani and Edouard Geoffrois

Zygmunt Vetulani, Edouard Geoffrois, Wojciech Czarnecki and Bartłomiej Kochanowski, *Language Resources for Public Security Applications: Needs and Specificities*

14:20 – 15:00 – Invited Keynote Talk by Chris Cieri (University of Pennsylvania, USA, Linguistic Data Consortium, Executive Director), *Language Resources for Public Security Applications: a Data Center Perspective*

15:00 – 16:00 – Resources (oral presentation)

Adam Dąbrowski, Szymon Drgas, Paweł Pawłowski and Julian Balcerek, *Development of PUEPS - corpus of emergency telephone conversations*

Irina Temnikova and K. Bretonnel Cohen, *The Crisis Management Corpus and its Application to the Study of the Crisis Management Sub-language*

Christian Fluhr, Aurélie Rossi, Louise Boucheseche and Fadhela Kerdjoudj, *Extraction of information on activities of persons suspected of illegal activities from web open sources*

16:00 – 16:30 Coffee break

16:30 – 17:15 – Poster session

Carlo Aliprandi, Tomas By and Sérgio Paulo, *Language Processing and Linguistic Data in the CAPER Project*

Richard Beaufort, Alexander Panchenko and Cédrick Fairon, *Detection of Child Sexual Abuse Media on P2P Networks: Normalization and Classification of Associated Filenames*
Simona Cantarella, Carlo Ferigato and Evans Boateng Owusu, *Design of a Controlled Language for Critical Infrastructures Protection*

Ales Horak, Karel Pala and Jan Rygl, *Authorship Identification to Improve Public Security*
Wiesław Lubaszewski and Michał Korzycki, *Unexpected Factual Associations Mining*

Miriam R L Petruck and Gerard de Melo, *Precedes: A Semantic Relation in FrameNet*

Milan Rusko, Sakhia Darjaa, Marian Trnka, Miloš Cerňak, *Expressive speech synthesis database for emergent messages and warnings generation in critical situations*

Zygmunt Vetulani, *Language Resources in a Public Security Application with Text Understanding Competence. A Case Study: POLINT-112-SMS*

17:15 – 18:00 – General discussion (animated by Frédérique Segond)

Editors

Zygmunt Vetulani	Adam Mickiewicz University in Poznań, Poland
Edouard Geoffrois	Direction Générale de l'Armement, Mission for Scientific Research and Innovation (MRIS), Bagneux, France

Workshop Organizers/Organizing Committee

Zygmunt Vetulani	Adam Mickiewicz University in Poznań, Poland
Edouard Geoffrois	Direction Générale de l'Armement, Mission for Scientific Research and Innovation (MRIS), Bagneux, France

Workshop Programme Committee

Laura Chaubard	Direction Générale de l'Armement, DGA Ingénierie de Projets, Bagneux, France
Edouard Geoffrois	Direction Générale de l'Armement, Mission for Scientific Research and Innovation (MRIS), Bagneux, France
Jakub Gorczyński	Polish National Police, Poznań, Poland
Fryni Kakoyianni	University of Cyprus, Nicosia, Cyprus
Nasrullah Memon	University of Southern Denmark, Odense, Denmark
Mario Montoleone	Salerno University, Italy
Karel Pala	Masaryk University, Brno, Czech Republic
Frédérique Segond	Viseo, Grenoble, France
Tadeusz Tomaszewski	University of Warsaw, Poland
Zygmunt Vetulani	Adam Mickiewicz University in Poznań, Poland

Table of contents

Author Index

Preface/Introduction

Workshop Description

Public security in Europe and in the World is facing several threats. These include threats connected with intended human activities such as terrorism, spontaneous risks related to uncontrolled behavior of individuals involved in mass events, natural disasters, etc. Combating these dangers generates challenges for information and communication technologies which in many cases directly involve various forms of natural language processing. Gathering, maintaining and processing language resources specific for security applications is of primary importance for the language technologies concerned. In some cases it appears useful to investigate and use sensitive linguistic data which generates technological and legal problems connected with privacy, ownership, civic rightsprotection, etc.

The workshop is intended to serve as a thematic discussion forum open to:

- language resources suppliers,
- researchers and language engineers interested in the development of systems for security applications involving language technologies,
- potential/actual users of such systems,
- people concerned with legal aspects of gathering, maintenance and applications of language resources for public security purposes.

Generation of a long term cooperation projects involving the workshop participants would be adesired side effect of the workshop.

Areas of Interest

The workshop will focus on the knowledge serving applications serving public security. Particularemphasis will be given to the crucial role of language resources and related technologies. Contributions are invited on – but not limited to – the following topics:

- security specific corpora,
- security specific terminology,
- language models for specific sub-languages and language registers important for security research,
- language technology based tools to enhance public security,
- linguistic tools for risk assessment,
- controlled languages for public security applications,
- AI and NLP decision supporting systems,
- sharing and processing sensitive linguistic data,
- legal aspects of security-oriented natural language processing and engineering,
- access to sensitive data,
- IPR issues,
- protection and use of sensitive source data,
- international collaboration issues,
- issues related with national and international funding

Language Resources for Public Security Applications: Needs and Specificities

Zygmunt Vetulani[1], Edouard Geoffrois[2], Wojciech Czarnecki[1], Bartłomiej Kochanowski[1]
[1] Adam Mickiewicz University, Poznań, Poland
[2] DGA, Mission for Scientific Research and Innovation (MRIS), Bagneux, France
E-mail: vetulani@amu.edu.pl, edouard.geoffrois@dga.defense.gouv.fr, w.czarnecki@amu.edu.pl, bartkoch@amu.edu.pl

Abstract

Language technologies and the associated language resources necessary to develop them are needed in a number of applications in the public security sector, and there is a growing demand for such applications. The paper illustrates the scope and importance of the needs by presenting various examples of applications along with the corresponding language technologies and language resources. However, collecting and sharing these resources can be especially difficult in that sector due to its specificities. The paper proposes to better identify and acknowledge these specificities in order to better address them and suggests that sharing experience across the various applications within the sector might help to overcome the difficulties.

Keywords: language resources, public security

1. Introduction

Language technologies are useful in a number of applications in the public security sector, and the demand for such applications is growing. Indeed, public security in Europe and in the World is facing various challenges. These challenges involve threats connected with human activities such as terrorism, spontaneous risks related to uncontrolled behavior of individuals involved in mass events, technological or natural disasters, etc. Combating these dangers generates challenges for information and communication technologies which in many cases directly involve various forms of natural language processing.

In order to develop these technologies, language resources are needed, including resources specific to the concerned applications and technologies. However, due to the specificities of that sector, collecting and sharing such resources can be especially difficult, or at least perceived as such. Indeed, one might need to investigate and use sensitive linguistic data which generates technological and legal problems connected with privacy, ownership, civic rights protection, etc. This is an impediment to the development of the technologies, and the lack of adapted technologies in turn limits the interest for further technological developments.

In order to properly address these issues, they must be clearly identified and acknowledged. The present paper and the related workshop it introduces aim at prompting discussions in that direction. The following sections reviews public security applications related to language technologies and the specific issues with the associated language resources.

2. Typology of public security related problems

An overview of the typology of main public security related problems will help us to better understand the field from the point of view of information technology needs, especially those involving natural language processing. We take in consideration the following classifications.

• **With respect to protection beneficent (object)**
- humans (individuals, groups, public order)
- intellectual goods
- material goods, infrastructures

• **With respect to protection providers**
- governmental and public organizations (e.g. administration, army, police, self-governance organizations)
- private companies

• **With respect to nature of threats**
- organized crime
- terrorist activities
- technological threats (e.g. severe accidents)
- natural disasters[1]
- pandemics
- threats to public order

• **With respect to territorial extent of threats**
- global
- regional
- local
- space irrelevant

• **With respect to temporal extent of threats**
- instant
- permanent
- temporary
- time neutral

(We do not consider this list as exhaustive.)

[1] The portal Global Data Vault provides information on the main disaster risks for the US territory. The portal enumerates the following threats (and provides risk maps for the USA): earthquakes, floods, hurricane, lightenings, tornados, thunderstorms, tsunami, volcano eruptions, wildfires. (http://www.globaldatavault.com/natural-disaster-threat-maps.htm; last visited on March 22, 2012).

3. Technological aspects of threats

Half a century ago a new epoch in the history of homo sapiens had just started, and this new epoch, the Information Society Age, gave birth to an Information Society which is a novel kind of social structure where humans will be surrounded by a new generation of information-rich artifacts and technologies designed to interact and to collaborate with human users[2]. This situation creates new risks as the human-oriented information based technologies may also serve criminal purposes. Saturation of the human environment with high technological artifacts augments the risk of technological disasters[3] which, in extreme cases may combine with traditional natural disasters. Information technologies may, and should, enter the game[4]. As the human factor is decisive in almost all prevention, protection and rescue operations, and communication is often the weakest link of a chain, the human language related technologies come at the front line of the future security technologies.

4. Technologies and resources

Human Language Technologies are covered by EU policies since the 1980s under successive EU framework programs[5]. The program to build in Europe a strong and competitive language industry able to produce the Information Society infrastructure was explicit defined within the IST Program (Information Society Technologies, also alled User-friendly Information Society), during 1998-2002, under the 5FP.

The concept of language technologies has been evolving since the very beginning and the list of particular sectors has kept changing. Further in this study we will refer to the following list of HL Technologies[6] extracted from the final program of the LREC 2010. More precisely, we have considered the names of thematic sessions. This is expected to provide a relatively complete coverage of the field, i.e., to be representative of the research activities of the last few years.

- **Text technologies**
  Anaphora, Co-reference (detection & annotation)
  Authoring Tools and Text Analysis
  Dialogue Annotation
  Discourse Annotation
  Emotion Detection and Expression
  Parsing
  Information Extraction
  Information Retrieval
  Knowledge Discovery
  Machine Translation
  Morphology Processing
  Natural Language Generation
  Named Entity Recognition
  Opinion Mining and Emotions
  Part-of-Speech Tagging
  Question Answering
  Semantic Annotation
  Semantics
  Syntax
  Text Dialogue Management
  Temporal and Spatial Annotation
  Understanding
  Word Sense Disambiguation and Evaluation
- **Speech technologies**
  Pronunciation Interpretation
  Speaker Identification
  Speech Dialogue Management
  Speech Processing
  Speech-To-Text
  Text-To-Speech
  Voice Synthesis
- **Non-verbal technologies**
  Body Language
  Emotion Expression
  Sign Language
- **Web technologies**

The necessary condition in order to build and to develop language industries is the availability of language resources necessary to develop applications which meets the Information Society needs and expectations (language technology applications in the field of public security do not make exception to the rule). The concept of language resources (LR) was "invented" and promoted by the visionary pioneer of language industries Antonio Zampolli. Zampolli (1996) defined this concept as meaning "written or spoken corpora, lexical data bases, grammars". For the purpose of this paper we established the typology of language resources in similar way as we did for language technologies.

- **Text resources**
  Text Dialogue Corpora
  Discourse Annotation Corpora
  Grammars
  Lexical Resources
  Lexicon Grammars
  Lexicons
  Morphological Resources
  Morphologically Annotated Corpora

[2] Cf. (Vetulani & Uszkoreit, 2011), preface.
[3] E.g. the tsunami in Fukushima nuclear plants in 2011.
[4] From the main page of the Polish Platform for Homeland Security portal (http://www.ppbw.pl/ppbw/en-ppbw.html; last access March 22, 2012): ”Public security in Europe is facing a host of serious threats. These include electronic crime, terrorism, cyberterrorism, organized crime (both criminal and economic) and drugs-related crime. Combating these threats creates new challenges for law enforcement institutions, with respect to both preventing and fighting these phenomena. Effectiveness of the institutions that are responsible for security depends more and more on their use of new technologies and computer-based solutions.”
[5] http://cordis.europa.eu/fp7/ict/language-technologies/ (last access Mars 22, 2012)
[6] The following lists of technologies and resources do not pretend to be exhaustive

Multimodal Annotated Corpora
Multiword Expressions and Collocations
Named Entities
Temporal Expressions (Corpora)
Terminology
Text Corpora
Treebanks
Web Resources
Wordnets and Ontologies

- **Speech Resources**
  Broadcast News
  Disordered Speech Corpora
  Speech Corpora
  Speech Data
  Speech Dialogue Corpora
  Text-to-Speech Corpora
- **Hybryd Resources and Tools**
  LR Infrastructures

5. Application examples and their resources

The public security sector is stigmatized by permanent rivalry between “good” ad “bad” technology use (the bad one being often a bit in advance). This rivalry, as well as the tradition of human active resistance in face of natural disasters is at the origin of a number of projects based on computer technologies intended to help combating the threats.

1 Threats for humans (individuals, groups, public order)
1.1 Detection of internet-based sexual criminality (pedophilia detection)
1.2 Internet crawling for terrorist activities detection
1.3 Preventive monitoring of mass events (e.g. football match, concert, demonstration)
1.4 Public information about current natural disasters (like flood, tornado, hurricane, volcanic eruption, earthquake, heatwave, landslide, etc.)**;** monitoring and coordination of rescue actions at natural or technological disasters (fire, flood, ..., train crash, nuclear plant explosion etc.)
2 Threats for intellectual goods
2.1 Cultural heritage objects protection and monitoring (antiterrorist)
2.2 Detection of large scale plagiarism
3 Threats for material goods and infrastructures
3.1 Antiterrorist monitoring of public infrastructure; flight / marine / metro / railway security monitoring

Fig. 1. Selected application types

As information is one of key issues in the security oriented activities, many of the application projects use human language technologies. It is well known since a long time that real-size applications which involve intensive natural language processing usually depend on real size language resources (grammars, dictionaries, ontologies, etc.) and that – for many languages – appropriate resources are hardly accessible or simply non-existent.

In this section we will review a list of selected application types, provide examples and characterize necessary language resources[7].

In what follows we will provide seme examples and comment on the application types listed above.

5.1.1. Detection of internet-based sexual criminality

Detection of cybercriminality, especially in the area of pedophilia, has attracted multiple researchers because of great social and political importance of the problem. Internet criminality has a worldwide impact and it is difficult to fight against it. It uses AI methods as e.g. language understanding based chatting systems.

Already in 1990s tools have been proposed[8] for finding suitable chat groups based on user-provided parameters. This software can perform chat room topics detection based on users' messages and as such can be used for initial selection of chat rooms that should be considered for future investigation.

ChatTrack (Bengel, J., et al. 2004)[9] is a tool developed at the University of Kansas[10] for a chat topic identification (e.g. on IRC or chat websites). It may be useful for supporting crimes detection and prevention. Application to online pedophile detection is possible after providing it with special dictionaries and corpora (for training it to identify adult and child language), so it can recognize adults posing as children.

Discovering the pedophilia chat servers requires, besides the "usual" resources necessary for development of language understanding systems, also specialized ones: corpora of chatting logs, specialized dictionaries, frequency lists for both general language corpora and for special corpora of criminal chats. Gathering experimental

[7] A number of important applications (tools) specified at a more general level are missing from this list. These are operational or forensic tools for helping e.g. speaker identification, creation of a psychological model of language user, monitoring of telephone calls, consistency check of witnesses' statements.

[8] E.g. by Butterfly Software (Van Dyke et al. 1998), according April Kontostathis et al. in: http://webpages.ursinus.edu/akontostathis/TextMining2009BookChapter.pdf

[9] May be found at: http://fciencias.udistrital.edu.co:8080/documents/12691/29360/10.1.1.10.3185.pdf.

[10] http://www.ittc.ku.edu/techtransfer/documents/ChatTrack.pdf

data, as well as using operational techniques as provocation, although technically possible may is difficult and in most cases creates legal and ethical problems.

5.1.2. Internet crawling for terrorist activities detection

Internet crawling for anti-terrorist and anti-criminal purposes is already widely used around the world. The general purpose is to identify traces of terrorist/criminal activities, at the earliest possible stage, as well as to discover of criminal organizational structures and interrelations and to identify individuals and their roles. Carnivore (US Government Surveillance of Internet Transmissions)[11] , implemented by the FBI is an example of a system to monitor email and electronic communications able to perform semantic analysis. This system, operational since 1997 was replaced in 2005 by a more performant Narus Insight Intercept Suite (NIS)[12]. In order to provide such systems with language competence at the deep understanding level it is necessary to use, besides basic resources as required for deep understanding systems, huge corpora representative of the application domain as well as specialized ontologies and lexicons.

DarkWeb Project[13] developed by Artificial Intelligence Laboratory of The University of Arizona is currently, for the authors best knowledge, the largest open-source corpus of terrorists and extremists forums, websites and blogs. Besides huge (over 500,000,000 sites/files/postings) number of resources, researchers developed a number of specialized tools like authorship identification system based on both syntactic and semantic analysis or clustering algorithms (e.g. Blockmodeling) used for terrorists social network analysis.

5.1.3. Preventive monitoring of mass events (e.g. football match, concert, demonstration)

Mass events are considered as being exposed to uncontrolled risks of public order distortion. Such phenomena use to be unpredictable. As human life is often endangered once the situation gets out of control, early prevention is extremely important. The prevention would consists here in the correct anticipation of what is going to happen on the basis of monitoring of the scene by many agents active in various places but able to communicate and to exchange their partial knowledge. The role of the information processing system is to assist the diagnostic process on the basis provided by the human agents in natural language. Both written and spoken forms may be considered as optimal, depending on the circumstances. The main technologies involved are those of language understanding and reasoning and the system must have characteristics of an intelligent system with natural language competence (multilingual in some cases) .

Dialogue Corpora
Grammars
Lexical Resources
Lexicon Grammars
Multiword Expressions, Collocations
Terminology
Text Corpora
Wordnets and Ontologies
(Speech Corpora
Speech Data
Text-to-Speech Corpora
Speech-to-Text)

Fig. 2. Useful resource types for language understanding AI systems

An implemented example of such system is POLINT-112-SMS developed at the Adam Mickiewicz University. The prototype of this system was tested at high risk football matches with participation of a large number of football fans. The system has a functionality of an expert system interacting with security staff in form of short text messages (SMS). (Vetulani et al. 2010; Vetulani & Marciniak, 2011).[14]

5.1.4. Public information about current natural disasters; monitoring and coordination of rescue actions at natural or technological disasters

The two main life-saving-functions of systems for public information about disasters is awareness and early alert in case of an approaching event. If the first function has a preventive character and results with better preparation to the critical moment, the second one is to trigger the immediate and appropriate reaction. For several reasons the warning messages should have a multimodal and often multilingual character which implies that various technologies should be involved: speech generation, sound, text, picture, animation. In some cases, the concerned area is huge and the propagation speed is high (tsunami, earthquake, forest fire, ...). In some cases the deed-back communication between the target addressees and the system will be necessary.

Global Disaster Alert and Coordination System (GDACS) is a "cooperation framework between the United Nations, the European Commission and disaster managers worldwide" for alerting about major sudden-onset disasters (like hurricanes, earthquakes, etc.). It started in 2004. It collects and organizes various data types, from GIS data (in-situ sensor data, satellite images) to media (and social media) information for some 14,000 disaster managers in the world. It is not clear if the process is

[11] http://www.vjolt.net/vol6/issue2/v6i2-a10-Jennings.html
[12] http://www.narus.com/index.php/product
[13] http://ai.arizona.edu/research/terror/
[14] http://www.springerlink.com/content/dxm42n615449uw74

currently automated. [15]

The Twitter Earthquake Detection system aims at finding, collecting, analyzing and visualizing information about earthquakes world-wide, making use of the most popular microblogging platform – Twitter. Application uses simple earthquake-related keywords dictionary to correctly identify tweets (Twitter posts) related to this kind of danger and performs extraction of the important data (like earthquake location, its strength etc.) which are later visualized on the world map.[16]

Typical language resources for public information systems are those fundamental for (multilingual) speech synthesis and text generation.

Monitoring and coordination of rescue actions in case of natural disasters (5.1.4), especially on a large area will require intensive communication within the organization responsible for this action. Therefore the necessary technological support will be similar as in case of monitoring of mass events (see the "ad 5.2.1", below).

5.2.1. Cultural heritage objects protection and monitoring (anti-terrorist)

Threats for cultural heritage are due to criminal or terrorist activities and are dealt by of security services both at the governmental and local level. Although the antiterrorist protection of the cultural infrastructure (museums, historical monuments of exceptional value) share most of characteristics with the protection of other kinds of terrorism-attracting installations there are some important differences as the permanent presence of multilingual crowd, many children around, etc. Protection against criminal acts, as theft or commercialization of stolen works of art constitute other kind of challenge. Some activities have been already observed on this field. E.g. the INTERPOL General Secretariat together with the Italian Carabinieri led a project PSYCHE (Protecting System for Cultural HEritage) [17] aiming at modernizing the INTERPOL's Stolen Works of Art database.

5.2.2. Detection of large scale plagiarism

On the one hand, widespread access to world wide web has made it much easier to plagiarize others work, but on the other hand – it has given a powerful tool for fighting such crimes. Plagiarisms take many different forms, from simply signing someone else work with your own name, through stitching together different works (without actual original input), to even self-plagiarism, which involves resubmitting your own work with only minor (or none) changes in its content. There are numerous examples of existing projects concerning this issue, some of which are listed below.

Plagiat.pl is the biggest plagiarism detection system used in Poland, deployed on over 80 higher education facilities. It serves both institutions that want to ensure that their students/employees work is original, as well as private users willing to check if their work is free of some legal violations.

Viper[18] - the Anti-plagiarism Scanner – is a free tool developed by Angel Business Limited for comparing documents against the database of over 10,000,000,000 sources or by user selected files stored on one's own hard drive. It also provides the user with interactive side-by-side comparison of uploaded document with detected (possibly) plagiarized one(s).

Useful resources:

Large text corpora (e.g. books or scientific papers)
Lexicons of common words and phrases that should not be considered as plagiarism.

Useful technologies:

Web crawling
Language modeling

5.3.1. Antiterrorist monitoring of public infrastructure; flight/marine/metro/railway security monitoring

We imagine application of closed-circuit television (CCTV) systems to register objects and events on the restricted area in a limited time. Instead of relying on human supervision only, the system should enable automatic image processing (exploration) in order to obtain a spatio-temporal map of the site. The system should involve an NLP module to interact with human analysts in order to validate system's findings, contradictions, introduce additional information etc. For this purpose, the system could be combined with systems such as the ones described in section 5.1.3. (e.g. POLINT-112-SMS). Systems partially responding to this challenge do exist or are under construction, especially for airport security, for example Nextiva IP Video Solutions for Airport Security[19] (an already operational system of the US Department of Homeland Security) aiming at Complex visual surveillance system for airport security containing modules for automated detection of e.g. perimeter intrusion, excessive vehicle speed (or wrong direction), objects left behind or camera tampering. Currently the project does not support natural language processing communication. In Europe, similar objectives were addressed by the "Surveillance of Unattended

[15] http://www.gdacs.org/

[16] http://ecopolitology.org/2010/01/07/usgs-develops-twitter-based-earthquake-detection-system/

[17] http://interpolnoticeremoval.com/tag/protecting-system-for-cultural-heritage/

[18] http://www.scanmyessay.com/

[19] http://verint.com/video_solutions/section2a.cfm?article_level2_category_id=2&article_level2a_id=375

Baggage and the Identification and Tracking of the Owner (SUBITO)" - a project funded by the 7 FP (2009-2011) [20] and the currently running project "Total Airport Security System (TASS)"[21] (until May 2014).

Useful resources: as those of Fig.2. (text).

Specialized project own resources may also be useful like: annotated corpora of image descriptions, experiment-driven annotated scene recordings/pictures, as well as projects specific data in form of regulations, scenarios and procedures (these last category of data has often confidential character and requires special care).

6. Conclusions

Language resources are necessary in public security oriented application as they are in all other kinds of applications and present a number of usual technological problems (such as acquisition problems for speech data due to noisy environment, non-representativeness of available corpora, etc.). There exist however problems so specific and challenging that language resources for public security applications are to be considered as a well-defined sub-sector of the area of language resources.

One of the reasons for this is a sensitive nature of the necessary linguistic data which generates technological, ethical and legal problems connected with privacy, ownership civic rights protection etc. The reader will find below a list of examples of particular threats and problems.

1. Problems with data (corpora) availability
1.1 Resources exist, but are private
1.2 There are no resources for one language, but there are for other(s)
2. Problems with obtaining corpora
2.1 Corpora need to be acquired from the groups (environments), against which the developed application is targeted (e.g. terrorism)
2.2 Obtaining some corpora require the appearance of particular accidental causes (e.g. natural disasters)
3. Ethical problems
3.1 Obtaining corpora in order to develop provocation of child abusers
4. Legal problems
4.1 Protection of data
4.2 The right to privacy (Privacy Policy)
4.3 Cases in which it is prohibited to disseminate (disclosure) of data (corpora) received due to protected nature of the sources (e.g. police records)
5. Problem with evaluation and representativeness of data
5.1 How to assess the representativeness of data collected during the experiments, when we do not have corpora to compare and check for correctness?
6. The instability of models
6.1 Some of the issues considered here are very dynamic (unique) in its nature (e.g. every catastrophe is quite different with respect to any other), so once prepared resource, system or method may fail in the next occurrence - there is a need to create systems and methods to dynamically "update" the model

We do not consider this list as exhaustive. Rather we want to bring the attention to the existence of a complex groups of problems that we are facing while acquiring resources and developing applications in the field of public security. We propose an open discussion of the domain experts in order to see which of the above identified threats and problems are the most important for the development of the sector.

What we propose as the main questions for further discussion are:
"Which other issues related to language resources for public security can be identified?"
"Which of those issues are the most challenging?"
"How to deal with them?"
"How to improve the acquisition of the necessary resources for public security applications?"

7. References

Bengel, J., et al., ChatTrack (2004). Chat Room Topic Detection Using Classification. *Lecture Notes in Computer Science*, Vol. 3073: pp. 266-277.

VanDyke, N. W., Lieberman, H. and Macs, P. Butterfly (1998). A Conversation-Finding Agent for Internet Relay Chat. *Proceedings of the 4th international conference on Intelligent user interfaces*, Los Angeles, California, USA. ACM Press, New York, pp. 39 – 41

Vetulani, Z. and Uszkoreit, H. (eds.) (2009). *Human Language Technology. Challenges of the Information Society. Third Language and Technology Conference, LTC 2007. Poznań, Poland, October 2007, Revised Selecterd Papers*, LNAI 5603, Springer.

Vetulani, Z., Marcinak, J., Obrębski,J., Vetulani, G., Dabrowski, A., Kubis, M., Osiński, J., Walkowska, J., Kubacki, P., Witalewski, K. (2010): *Zasoby językowe i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego (in Polish)* (Language resources and text processing technologies. POLINT-112-SMS as example of homeland security oriented application), Adam Mickiewicz University Press: Poznań.

Vetulani, Z., Marcinak, J. (2011). Natural Language Based Communication between Human Users and the Emergency Center: POLINT-112-SMS. In. Z. Vetulani (ed.), Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2009. Revised selected papers. LNAI 6562, Springer, pp. 303-314.

Zampolli A., (1996). Współpraca międzynarodowa w dziedzinie LR (in Polish), Informatyka, Nr 3, (English title: International co-operation in the domain of Language Resources), pp. 34-37.

[20] http://www.subito-project.eu/

[21] http://www.tass-project.eu/

Language Resources for Public Security Applications: a Data Center Perspective

Chris Cieri

University of Pennsylvania, USA, Linguistic Data Consortium

Abstract

Among the many corpora that LDC is producing or distributing, several, for example some of the Mixer corpora, are related to public security variously defined. In this talk we present some of these corpora and how they were created. We also describe some of the issues encountered in their creation which are related to the public security domain, how we overcame them and the lessons learned. Some specific issues we will discuss include matching data specifications to rapidly evolving requirements, managing intellectual property, protecting the privacy of human subjects and distributing resulting data.

Development of PUEPS corpus of emergency telephone conversations

Adam Dabrowski, Szymon Drgas, Pawel Pawlowski, Julian Balcerek

Chair of Control and Systems Engineering, Poznan University of Technology
Piotrowo 3A, 60-965 Poznan
{Adam.Dabrowski,Szymon.Drgas,Pawel.Pawlowski,Julian.Balcerek}@put.poznan.pl

Abstract

In this article development of a PUEPS corpus is described. This dataset contains recordings of the acted emergency telephone conversations. Speakers that participated in the experiments reported crime scenes that were presented to them in a form the earlier prepared movies. Recording sessions were performed in the laboratory conditions. To each conversation metadata that summarize information about the speaker, conversation, and the reported event were added. Moreover, manually prepared transcriptions enriched with tags describing paralinguistic phenomena are also a part of the described corpus. These transcriptions were made using tools prepared by the authors for fast and convenient work due to: prompting, annotation, and data management mechanisms. The transcription experiments showed substantial improvement of the work efficiency and speed. Final multilevel speaker recognition experiments proved that the accuracy of the speaker recognition is noticeably improved due to the use of transcriptions and the linguistic level analysis.

Keywords: transcription, metadata, speaker recognition

1. Introduction

Emergency telephones are often misused by persons who cause false alarms. Such incidents are expensive and potentially dangerous (for example in case of an unnecessary evacuation of a hospital). In order to avoid improper uses of emergency telephones, there is a need for monitoring of the callers. This can be realized by running a specially prepared system for fast speakers classification and their identity recognition.

In order to develop such a system and evaluate methods together with respective algorithms for the automatic classification of the stored conversations there is a need for preparation of a corpus of relevant test conversations. A possibility of obtaining such data from the Police does not work in practice due to legislative restrictions and difficulties of gaining more than one conversation from the same speaker in a prescribed period. Therefore we have prepared acted but certainly quasi real and relevantly balanced data. They are vulnerable, anonymous, and thus do not need any special care.

There are several characteristic features of emergency telephone conversations. In most of them there are two speakers talking only. Emergency telephone operator and a person who is reporting an incident that he/she is a witness of or a victim. The emergency telephone operator tries to posses information about the calling person (name and surname), about the incident (location, time, what happened, etc.), about people that took part in the event (number of the people, their characteristics), perpetrator (direction where he went), etc. The average time of conversation is about 1 minute. The quality of recordings is distorted by the telephone communication channel. In most cases additional undesired sound sources (acoustic background) are also recorded.

Although, some corpora with audio data for Polish language were already prepared, there exists no corpus that would be appropriate for the speaker recognition experiments for emergency telephone conversations. the data base reported in (Grocholewski, 1995) is a corpus annotated at the phoneme level. It is recorded in laboratory conditions and it consists of read sentences, names, and numbers only. A newer example is the data base Jurisdic (Demenko et al., 2008). It was designed for a special application of transcription of utterances in court. It consists of read speech and semi-spontaneous answers to questions. The annotations are available at the word level only. Another corpus that can be mentioned is Luna (Raymond et al., 2007). It includes human-machine dialogs in Italian, Polish, and French. It contains annotations at syntactic, semantic, and discourse levels.

Beside Polish corpora among other data sets with recordings prepared specially for the speaker recognition tests are various versions of Mixer used in the NIST speaker recognition evaluations (Przybocki et al., 2007). They contain hundreds of telephone quality dialogs.

As it was already mentioned the authors decided to record acted conversations in the laboratory conditions. There are several advantages of such approach. First, there is a possibility to control a degradation of the signal (it can be mixed with acoustic background recordings and it can be distorted by a telephone communication channel). Moreover, there is a possibility to obtain many recordings of the same person, the same person reporting the same incident, or different persons reporting the same incident.

In this paper, it is presented how to collect and process acted emergency telephone conversations. Our corpus contains the following kinds of data:

- high quality recordings (laboratory conditions, sound insulated booth and high quality microphone),
- telephone quality recordings,
- manually prepared transcriptions that contain paralinguistic phenomena,
- metadata introduced by the telephone operator.

The paper is organized as follows: data collection process is described in Section 2.. Next, devices and connections needed to obtain data are presented in Section 3.. Then, the recording sessions report is provided in Section 4.. It is followed by the discussion about the metadata. Finally, the transcription process is described, typical applications are discussed, and the conclusions are given.

2. Data collection process

The data collection process consists of the following steps:

1. presentation of the movie with the crime scene to report,
2. telephone conversation, i.e. reporting the crime scene,
3. entering the metadata about the conversation by the telephone operator,
4. transcription.

In the presented corpus, it was very important to capture speaker dependent linguistic features. In order to avoid any suggestion of words, e.g., by giving verbal instructions to the speaker, silent movies with crime scenes had been prepared. Before the conversation, the movie was presented to the speaker. The task of the persons that took part in the recording session was to describe the situation from the movie and to answer questions asked by the operator. The persons that was playing a role of the operator were sitting in another room. Thus there was no visual contact between the person that reported the incident and the operator.

After presentation of the movie, the speaker had to establish telephone connection with the operator. The speaker was located in a sound insulated booth. The voice of the calling person was recorded twofold: using high-quality condenser microphone, and the signal directly from the telephone line.

After the conversation the telephone operator entered the metadata to the database. This was done by filling a standard form. The questions in the form were related to relevant information about the conversation. The fields in the form were designed in a way that helped to provide most reliable information as quickly as possible.

Finally, manually transcriptions ware performed. Transcribers were asked to do annotations aligned at the "phrase" level. Additionally, tags that describe paralinguistic phenomena were added. The tools for management are transcription are described in detail in Section 6..

Additionally, it is possible to add some earlier recorded acoustic backgrounds using the head and torso equipment (Cetnarowicz et al., 2010).

3. Apparatus

In order to achieve a reliable and an easy to manage process of the data collection and transcription a network based, distributed system was prepared. An architecture of this system is presented in Fig. 1. Particular processes were performed in several places, i.e. in an anechoic room, in the emergency phone operator office, and transcription offices.

In the anechoic room, equipped with the terminal, telephone and the high quality microphone, a caller can watch video sequences and make calls. The video playback is performed by the video server, then the caller realizes the call. A voice of the caller was recorded using high-quality microphones and audio recorders together with standard telephone quality recorders.

In parallel, the call (i.e. voices of both interlocutors together) were digitally stored with a recorder connected to the PBX (private branch exchange). The second recording was made with the telephone line quality. Additionally, a voice of the operator was watermarked with an extra made telephone adapter. This operator telephone adapter consists of a DSP (digital signal processor), an analog line interface, analog-to-digital and digital-to-analog converters. The audio watermark is almost unhearable and strongly helps to automatically separate the voices without any speaker recognition algorithm (Chmielewska et al., 2011).

During the call the operator prepares and stores relevant text data using a special dialog form and puts them into the database. The database is managed by the database server that stores the records with the calls. Each record consists of metadata either entered by the operator or produced automatically in the post-processing phase. Finally, the waveforms are transferred from the audio recorder and the call recorder.

The communication server synchronizes both the audio and the call recorder with the database server via LAN by means of the TCP/IP network stack.

After the recording stage, the calls have to be transcribed. To speedup and make this process easier an additional transcription server equipped with the web-based software was prepared. The transcriptions can be performed from any location, using typical computers equipped with web browsers only (see Section 6.).

The recording and the transcription system is based on the client-server architecture, which uses a variety of connections and standards to exchange data. There are five servers: communication, database, video, audio recorder, and the call recorder. On the client side there are also several types of interfaces: the caller that can see the video and enter some ID, the operator that manages the call description, the administrator of transcription process, and finally the transcript writer him(her)self. All clients have dedicated GUIs (graphical user interfaces). The user and communication interfaces as well as the software modules were prepared using C, scripts, HTML, SQL, PHP, and Java languages and ActiveX plugins.

4. Recording

A process of collecting the recordings was divided into sessions. Each speaker had to take part in at least 6 conversations. However, maximally two conversations for one speaker during one session (one day) were allowed.

There were 30 speakers (students, aged 19–25 years) that took part in the recording sessions, and four persons that took part of the telephone operator service.

As mentioned before the speakers were asked to report the crime scene seen before in the movie. They were instructed that when they will be asked about their personal data (name, address) they should give fictitious but stable

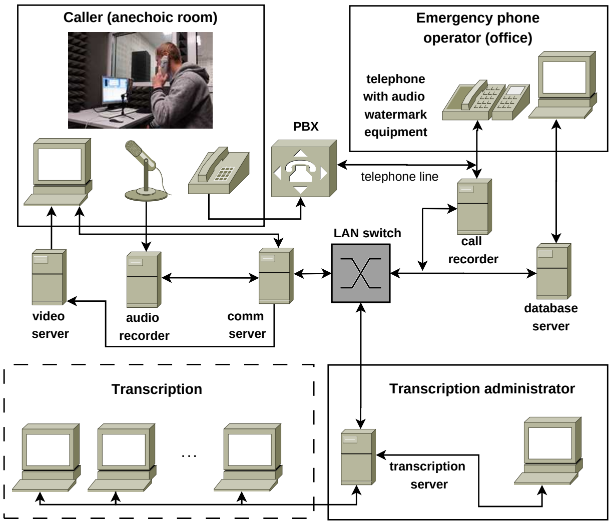


Figure 1: Recording and transcription system architecture

information. Persons who acted roles of the telephone operator had the task to gain the following information:

1. type of event (intervention, traffic incident, crime),
2. place of event,
3. time of event,
4. circumstances of the incident,
5. what are the effects of the incident,
6. is the perpetrator known? - his/her characteristic features, location, in case of escape — direction and type of the movement,
7. information about the reporting person.

Finally, 180 conversations were recorded. A mean time of the conversation is 1 minute.

5. Metadata

Generally, metadata may describe a lot of data types like textual data, audio files, images, video, or even 3D video data. In real world situations there is a need to select and combine relevant information from many sources of information in order to better achieve specified purposes. Metadata as a textual description or numbers are easy to be stored in the database because they have much smaller size than other data types like e.g. huge video files. Metadata describe characteristic features of objects, events, or situations. They can also describe features, which can change in time. The prepared database searching mechanisms operate on the metadata.

Chosen metadata, which are filled and processed, have an influence on the proper operator reaction and the results of searching. In order to find an adequate person or an event the metadata must clearly describe and distinguish relevant data.

During the call the operator asks a series of questions in order to obtain the necessary information. If it is available, the video information about the event from the monitoring system cameras may be used as an additional aid to the operator. We prepared a stereovision system, which can help the operator to focus his/her attention on the interesting image regions.

In our case the metadata carry information about the event, audio files, and person who called. For example: a kind and category of the event, name and surname, age, address of the caller, date and time, characteristic features of the voice, acoustic background, etc. A hint system embedded in the graphic user interface proposes proper metadata values for some metadata like an address or the date. Prepared selection lists are also included.

If we want to find an object in the database using metadata we have to compare a value of the searched item with values of the items stored in the database.

In the simplest case we can only determine if a value of a particular feature occurs in the database records or if it does not occur. As a result of such comparison, a binary '0' or '1' value is assigned. For non-numerical values it is possible to specify functions of the similarity between the searched data and the data stored in the database.

Values of such functions are stored in similarity matrices. A similarity matrix, which corresponds to the metadata, has the highest entries for exactly the same feature values. These arguments are placed on the main diagonal of the similarity matrix. For numerical values we can determine ranges of values for features and the corresponding weight values.

It is also possible to determine changes of parameter according to changes of other parameters. Such correlations may define changes of parameters in time and may be stored in the database also as the correlation matrix. All types of metadata comparisons are also weighted according to the importance of particular features. Our mathematical formulas for the database searching mechanism based on various types of metadata are described in (Balcerek et al., 2009), (Balcerek et al., 2010) and (Balcerek et al., 2011).

The proposed metadata searching mechanism may be used in a combination with other methods. For example, metadata searching mechanism is used to choose a list of similar objects and, then, this result can be adjusted precisely, e.g., by using the automatic voice analysis in case of audio files. Based on the information, which comes from the metadata, it is even possible to automatically control devices or to start some security procedures like sending an ambulance.

Metadata searching mechanism may also be adopted to other real world situations. It may be used for searching for a suspect with the gathered information. Another application can be a database of birds for ornithologists where the information about the birds can influence activities related to the nature conservation.

6. Transcription

After the recording stage, the speech material was – as already mentioned – manually transcribed. The authors prepared a dedicated software to perform both administration for assigning the transcriptions to given transcribe writers, and the transcription process itself. The software for administration of the transcriptions offers:

- management of users accounts,

- assigning new transcriptions to the workers (transcribe writers),
- edition and canceling the existing transcriptions,
- reading records from the database of calls,
- sending messages between users,
- storing of prepared transcriptions in the database for further processing.

After the transcription was assigned to the user he or she has to log in to the system using a PC equipped with a sound card and a web-browser (no additional plug-ins or installs are required). A location of the user is not limited, although it can be done by the administrative and network routing rules. The user interface for the transcription process is presented in Fig. 2. In the top of the screen the player is placed, in the middle a dialog is visible, and on the bottom the layer of the transcription is presented in a form of the the waveform plot as a function of time.

The transcriber marks borders of each phrase and additionally enters annotations enriched with tags to mark various phenomena that occurred during the utterance. A list of tags is presented in Table 6. Abbreviations came from the Polish language.

There are many features that improve simplicity, reliability, and efficiency of the work. Among them are:

- dictionary based contextual clues, with the whole phrases (not only single words),
- keyboard short-cuts, e.g. for control of the player,
- possibility of listening the call during typing,
- playing from the pointed and selected time,
- sending messages between users,
- storing of prepared transcriptions int the database for further processing.

Because the corpus consists of emergency telephone conversations, where some phrases are very common, it was expedient to prepare a dialog window with contextual clues. The prompted phrases, not only words, come from the database adaptively using the already prepared transcriptions. A speedup of the transcription process, that is a result of all proposed mechanisms reported above, as a function of the speed of the standard key typing is shown in Fig. 3. It can be noticed that even for less experienced users, the speedup reaches about 50%.

More than 180 calls were transcribed by 6 persons. In average, a transcription of a 1.5 minute dialog takes the advanced transcript writer more than ten minutes. A total time of all phrases after the transcription is equal to ca. 3 hours.

7. Applications of the corpus

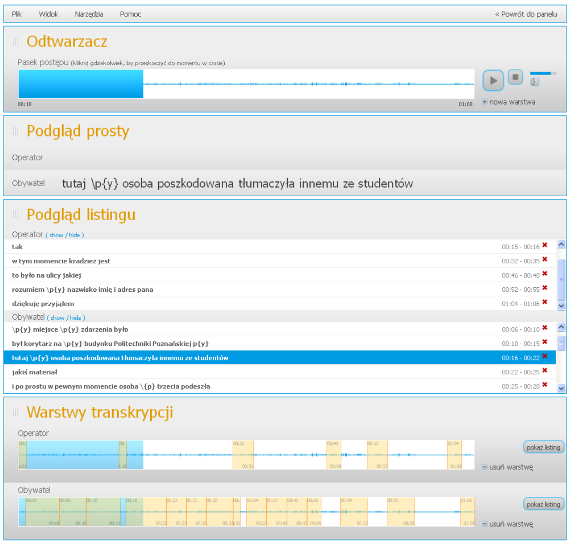
Although the corpus was designed for the experiments with the automatic speaker recognition based on the multilevel analysis of the speech signal, other applications are also possible. As the transcriptions with segmentation on the "phrase" level are provided, the corpus can be used for evaluation of automatic speech recognition systems for Polish language. It is also possible to use this corpus for the development of the speaker diarization algorithms.

An example of the experiment with the PUEPS corpus is recognition of speakers by means of multilevel speech signal analysis taken from (Drgas and Dabrowski, 2012).

Idiolectal aspects of speakers can be taken into account (Doddington, 2001). They were caught by lexical features obtained from the manually made transcriptions of the PUEPS corpus. First, dictionaries were constructed for the word bigrams. Then for each side the numbers of occurrences of all words in the dictionary were counted. These numbers were then used to construct bigram vectors for each side.

The results obtained for the speaker verification experiments are shown in Figure 4. There are plots of four curves. Beside the one that corresponds to the mentioned idiolectal (lexical) features there are also results obtained for spectral, prosodic, and articulatory features. Details of the feature extraction process can be found in (Drgas and Dabrowski, 2012).

It can be noticed that the EER obtained for the lexical bigrams is 33.50%. Although it is more than three times than



Operator

Obywatel tutaj \p{y} osoba poszkodowana tłumaczyła innemu ze studentów

Operator (show / hide)

tak	00:15 - 00:16
w tym momencie kradzież jest	00:32 - 00:35
to było na ulicy jakiej	00:46 - 00:48
rozumiem \p{y} nazwisko imię i adres pana	00:52 - 00:55
dziękuję przyjąłem	01:04 - 01:06

Obywatel (show / hide)

\p{y} miejsce \p{y} zdarzenia było	00:06 - 00:10
był korytarz na \p{y} budynku Politechniki Poznańskiej p{y}	00:10 - 00:15
tutaj \p{y} osoba poszkodowana tłumaczyła innemu ze studentów	00:16 - 00:22
jakiś materiał	00:22 - 00:25
i po prostu w pewnym momencie osoba \{p} trzecia podeszła	00:25 - 00:28

Operator

Obywatel

Figure 2: Tool for transcription

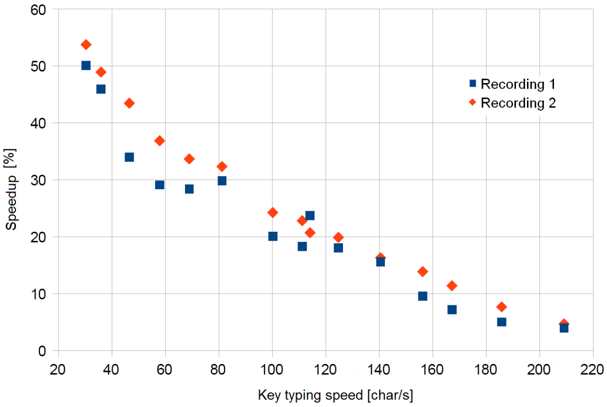


Figure 3: Speedup of transcription

Table 1: List of tags used in transcription

Sounds articulated by the speaker	
\p{}	Filled pause
\s	Laughing
\k	Scream
\p	Crying
\ka	Cough
\ch	Hawking
\zi	Yawning
Untypical pronunciation of existing words	
\a{}	Foreign accent
\n{intentional, trans.}	Incorrect pronunciation
\j{}	Stammering
\prz{}	Prolongation of a word
\k{}	Screamed words
\p{}	Words uttered during crying
\s{}	Words uttered during laughing
\nd{intentional, trans.}	Incomplete words
\f{}	Fast speech
\prr{}	Stopping words
Untypical words or phrases	
\g{}	Dialect expression
\z{}	Loan-words
\wu{}	Vulgarisms
\q{}	Abbreviations
\l{}	Lettering
\nd	Incomplete word
\f{}	Fast uttering
\mhm{yes/no}	Confirm./neg. by non-words
\w{trans.}	Unintelligible words

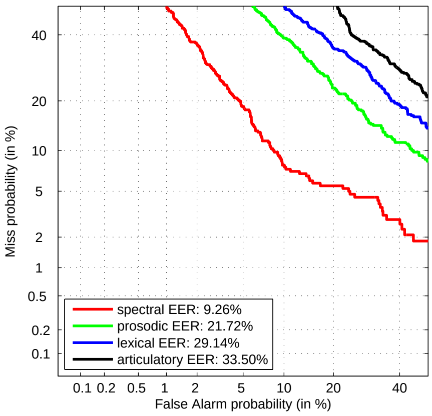


Figure 4: Results for SNR of 20dB

that for the for spectral features, this result indicates that the lexical features carry speaker dependent information. It can be complementary to the one provided by the spectral features, which is very important for the plausible speaker recognition.

8. Conclusions

In this paper a development of the PUEPS corpus has been reported in detail. The data collection process has also been described. Additionally, an efficiency of the mechanisms developed for the speedup of the transcription process has been quantified. Finally, an example of the application of the PUEPS corpus for the speaker verification experiments has been presented.

9. References

J. Balcerek, S. Drgas, A. Dabrowski, and A. Konieczka. 2009. Prototype multimedia database system for registration of emergency situations. In *Proceedings of Signal Processing SPA'2009, IEEE Poland Section Chapters Signal Processing, Circuits and Systems*, pages 144–148, Poznan, Poland, September 24-26.

J. Balcerek, P. Pawlowski, A. Konieczka, S. Drgas, A. Dabrowski, and M. Kmieciak. 2010. Database of emergency telephone calls - system tools for real-time registration and metadata searching. In *Proceedings of Signal Processing SPA'2010, IEEE Poland Section Chapters Signal Processing, Circuits and Systems*, pages 89–94, Poznan, Poland, September 23-25.

J. Balcerek, A. Dabrowski, S. Drgas, P. Pawlowski, and A. Konieczka. 2011. Database and recording system for registration, maintaining and fast searching of emergency telephone calls. *Elektronika - konstrukcje, technologie, zastosowania*, 5:116–123.

D. Cetnarowicz, S. Drgas, and A. Dabrowski. 2010. Speaker recognition system and experiments with head / torso simulator and telephone transmission. In *Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings (SPA), 2010*, pages 99 –103, sept.

A. Chmielewska, A. Dabrowski, A. Meyer, M. Portalski, and R. Weychan. 2011. Segmentation of speakers during emergency call with watermarking technique. In *IEEE SPA Conference 2011*, pages 106–111.

G. Demenko, S. Grocholewski, K. Klessa, J. Ogorkiewicz, A. Wagner, M. Lange, D. Sledzinski, and N. Cylwik. 2008. Lvcsr speech database - jurisdic. In *Signal Processing Algorithms, Architectures, Arrangements, and Applications (SPA), 2008*, pages 67 –72, sept.

G. Doddington. 2001. Speaker recognition based on idiolectal differences between speakers. In *Eurospeech-2001*, pages 2521–2524.

S. Drgas and A. Dabrowski. 2012. Speaker recognition based on multilevel speech signal analysis on polish corpus. In *Multimedia Communications, Services and Security, 5th International Conference, MCSS 2012*.

S. Grocholewski. 1995. Corpora - speech database for polish diphones. In *EUROSPEECH 95*.

M. A. Przybocki, A. F. Martin, and A. N. Le. 2007. Nist speaker recognition evaluations utilizing the mixer corpora 2004, 2005, 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):1951–1959, Sept.

Ch. Raymond, G. Riccardi, K. J. Rodriguez, and J. Wisniewska. 2007. The LUNA corpus an annotation scheme for multi-domain multi-lingual dialogue corpus. In *11th Workshop on the Semantics and Pragmatics of Dialogue*, pages 157–186.

The Crisis Management Corpus and its Application to the Study of the Crisis Management Sub-language

Irina Temnikova[1] and K. Bretonnel Cohen[2]

[1]Research Institute in Information
and Language Processing
University of Wolverhampton, UK

[2]Biomedical Text Mining Group
Computational Bioscience Program
University of Colorado School of Medicine
Denver, Colorado, USA

irina.temnikova@gmail.com, kevin.cohen@gmail.com

Abstract

This article presents a novel language resource, the Crisis Management Corpus (CMC). The corpus is the first in its domain and is expected to be of utility for linguistic studies and for natural language processing applications in the crisis management and the public security domains. The article describes the collection, pre-processing and composition of this resource, along with its possible applications. Two example applications of the resource are described in detail. The first application is the study of the text complexity levels characterizing the CMC, with the aim of evaluating the communicative efficiency of written documents in the domain. The second application is a preliminary investigation of the linguistic characteristics of the crisis management sub-language.

Keywords: language resources, crisis management, corpus linguistics

1. Introduction

Due to the fact that natural language can be very complex and ambiguous and that under stress conditions human comprehension is altered, e.g. because of very short reaction times (Kiwan et al., 1999; Ogrizek and Guillery, 1999; Winerman, 2009), several deadly disasters caused by human miscommunication have occurred, including the Tenerife air crash (Air Line Pilots Association, 1977) and the Scandinavian Star ferry disaster (Solheim et al., 1992). In addition, studies have shown that a large number of car accidents in the U.S.A. resulting in fatal injuries to children were due to incomprehensible instructions for child seat usage (DuBay, 2004). For these reasons, communication is an important factor for successful crisis management (CM) (Regester & Larkin, 2005) and one of the technological hazards (Coppola, 2007) that must be kept under control.

This article presents a novel language resource, the Crisis Management Corpus (CMC). It is the first corpus of its kind. It is a monolingual corpus in the English language. The corpus contains four sub-corpora of different types of CM documents, with the aim of ensuring representativeness of types of target audiences, Crisis Management sub-domains, and document types. The corpus is an essential tool for studying the language of the crisis management and public security domains and creating Natural Language Processing (NLP) applications for them. The corpus has already been used for three such tasks, namely:

1. A corpus analysis aiming to study the linguistic characteristics of the crisis management sub-language.
2. A text complexity corpus analysis in order to study the efficiency of communication in the CM domain, and
3. As a basis for the development of controlled language guidelines for writing clear CM and public security-related documents and rewriting existing complex documents into simple ones (Temnikova et al., 2012).

This article describes the Crisis Management Corpus, as well as the results of an analysis of its linguistic and text complexity.

The article is structured as follows. Section 2 presents the related work on NLP for the crisis management domain, Section 3 describes the corpus, Section 4 presents its possible applications, and Section 5 provides the conclusions.

2. Related work in natural language processing for the crisis management domain

Although several computer-based crisis management systems have been developed (for example Kienzle et al., 2010), the contribution of natural language processing (NLP) to the crisis management field and, in particular, to the public security field has been limited. Potential contributions of NLP can be split into approaches addressing the crisis detection stage and approaches addressing the communication management aspect.

Approaches addressing the detection stage consist of information extraction applied to detect emergency events on the web in sources such as Twitter, open discussion forums, blogs and online news articles (Corvey et al., 2010; Ireson, 2009; Steinberger et al., 2009), or to ensure epidemic surveillance via clinical notes (Conway et al., 2009) or newswire text (Doan et al., 2007). The approaches addressing communication management were the controlled languages guidelines tailored for written or spoken communication in the domain, such as AECMA (Unwalla, 2004) for the aeronautics domain, PoliceSpeak (Johnson et al., 1993) for communication in the Channel Tunnel, and LiSe (Renahy et al., 2010) for healthcare protocols. All of these aim to solve the communication problems described in Section 1 by restricting natural language complexity and ambiguity. To the knowledge of the authors, no crisis management corpora have previously been created, and no study has systematically, objectively, and on a large scale investigated the linguistic characteristics of communication in the CM domain.

3. Description of the Crisis Management Corpus

The Crisis Management Corpus (CMC) is a monolingual, finite, representative corpus tailored to the crisis management domain. The CMC was collected semi-automatically from the web (partially manually, partially automatically using the Firefox extension *Mozilla Scrapbook* [1]) and is composed of crisis management documents available to the general public.

In order to transform the corpus into machine-readable form, the documents were preprocessed with a series of Python scripts in order to transform.html and .pdf files into plain text and remove unnecessary elements. The whole corpus was also parsed with the dependency parser *Machinese Syntax* (Tapanainen and Järvinen, 1997). It provided part of speech taggin, syntactic structure, and syntactic dependencies. None of these were manually corrected.

The document sources include the U.S. Federal Emergency Management Agency (FEMA)[2], the U.S. Center for Disease Control and Prevention (CDC)[3], and the British Red Cross[4]. The composition of the corpus was dictated by the wish to have it be representative, with a variety of intended audiences (specialists and non-specialists), domains, and sub-languages (general crisis management, aeronautics, and medicine) and document types (instructions and alerts).

The resulting corpus consists of four sub-corpora:

1. Emergency instructions for non-specialist populations (GP)
2. Emergency protocols and emergency plans for crisis managers (Spec)
3. Emergency instructions for pilots[5](SC)
4. E-mail alerts of disease outbreaks, downloaded from http://www.promedmail.org/ [6](PMM)

Table 1 shows the size of the whole CMC and its sub-corpora, listing first the separate sub-corpora (marked from 1 to 4) and then the data for the CMC as a whole.

Sub-corpus	N. of files	N. of sentences	N. of words
1. GP	58	12 451	156 571
2. Spec	160	74 875	1 243 381
3. SC	44	21 511	299 175
4. PMM	1486	59 477	1 029 413
Entire CMC	**1748**	**168 314**	**2 728 540**

Table 1: Statistics for the entire CMC and its sub-corpora.

Table 1 shows the number of textual files in each corpus in the second column. This varies due to the different file sizes. The third column shows the number of sentences per corpus, and the fourth column shows the number of words. As can be seen, the largest sub-corpora are the sub-corpus containing protocols for crisis managers (Spec) and the corpus composed of ProMedMail medical alerts. The whole CMC has over 2 700 000 words.

4. Applications of the crisis management corpus

This section discusses some possible uses for CMC (Section 4.1) and illustrates two uses of it for studying the crisis management sub-language via corpus analysis (Sections 4.2 and 4.3).

4.1. Possible applications of the CMC

The Crisis Management field (CM) is developing at an exponential speed, and more and more NLP interest is turning towards this domain. For the moment there are almost no linguistic resources available for the domain. Since this is the first available CM corpus, it is likely that it will be useful for multiple types of NLP applications. As a primary contribution, the corpus can be used for studying the sub-language(s) of the CM domain or for training and evaluating machine learning approaches for the CM, as well as for testing and evaluation purposes. At this time, the CMC has been already used for three applications: the development of the Controlled Language for Crisis Management and its evaluations (Temnikova et al., 2012), and for two crisis management sub-language corpus analyses (presented in Sections 4.2 and 4.3).

Besides being applicable to the CM domain, the CMC is an addition to our stock of corpora containing

[1] http://amb.vis.ne.jp/mozilla/scrapbook/, last accessed on October 21st, 2011.

[2] http:// www.fema.gov/, last accessed on March 14th, 2012.

[3] http://www.cdc.gov/, last accessed on March 14th, 2012.

[4] http://www.redcross.org.uk, last accessed on March 14th, 2012.

[5] Downloaded from www.smartcockpit.com (last accessed on February 22nd, 2011).

[6] Last accessed on February 22nd, 2011.

instructions; these have been the source for valuable linguistic insights, such as the classic studies of ellipsis and anaphora in recipes (e.g. Kosseim et al., 1996).

Finally, from a corpus linguistics perspective, the CMC is an addition to the sets of registers and domains found in more general corpora, and can be used to increase the representativeness of stratified corpora such as the Brown corpus and its many more recent descendants.

4.2. Text complexity analysis of the CMC

Besides the application of the CMC for developing corpus-based controlled language guidelines (Temnikova et al., 2010), the CMC has been used to study the communication efficiency of crisis management documents. The next sub-sections will present the motivations leading to this analysis (Section 4.2.1) and the analysis itself (Section 4.2.2).

4.2.1. Motivations for the text complexity analysis

The initial motivation for conducting a text complexity analysis of CM texts was to assess whether crisis management documents are written in a register that is simple and clear enough to be able to deliver the transmitted message correctly. An additional motivation was to assess the need for developing a simplification approach for purposes such as enhancing comprehensibility by humans and translation by humans or computers.

According to psycholinguistic research on reading (Harley, 2008), there is a large amount of textual and language phenomena which can hinder reading comprehension. For purposes of clarity, they can be divided into lexical, syntactic, and discourse phenomena. Examples of lexical text complexity phenomena are ambiguous words, technical terms, and inconsistent terminology. Examples of syntactic text complexity phenomena are long sentences with convoluted syntax and ambiguous syntactic structures. Finally, examples of text complexity at the discourse level are illogical order of statements or unclear relationships between separate statements. The reasons why these issues hinder human comprehension is that if the reader encounters a text complexity or text ambiguity issue, she/he has to go back and re-check the meaning of the whole sentence to understand it. This can be very dangerous in an emergency situation, when there is no time to re-read the text.

In order to exemplify the need for systematic study of the text complexity phenomena affecting crisis management and public security documents, a text providing instructions for actions to take during terrorist attack has be examined.

The text comes from *Individual Preparedness and Response to Chemical, Radiological, Nuclear, and Biological Terrorist Attacks* (Davis et al., 2003), which is a thirty-five-page guide containing an analysis of strategies adopted and suggestions for actions to be taken during terrorist attacks. The text contains safety instructions for a non-specialist population to be executed during chemical attack.

The text is composed of a title and four actions to be executed, the first two offering actions in two alternative conditions and the second two in temporal order. It can be also seen that, although the document is designed for non-specialists to be read during an emergency situation and thus to needs to be straightforward to follow, the first two situations contain long and dense text, composed of many alternative conditions (*if.., and...*) and multiple actions (t*ake shelter... close windows... move upstairs...*). The order of actions and the separate situations are not shown clearly. Additionally, the information presented is composed of overly many convoluted informational units. If the text is analysed further for incidence of linguistic issues known to hinder human comprehension (Harley, 2008), it will be noted that it exhibits a large quantity of them, such as long sentences, complex syntax, too much information to remember, technical words, imprecisions, passive voice, and elisions.

4.2.2. Text complexity analysis results

The text complexity analysis of the CMC is based on the assumption that text complexity can be measured by using a number of computable measurements, on which all of the existing approaches to measuring text complexity are based.

Unlike the classic readability formulae (Flesch, 1948; Dale and Chall, 1948) and similarly to (McNamara et al., 2010), the text complexity analysis of the CMC did not attempt to give a unified score, but rather to investigate the quantitative presence of different text complexity issues in the CMC documents. The choice of the set of high TC issues analysed in the CMC was motivated by several reasons, among which were psycholinguistic findings about human comprehension under stress. In order to provide a baseline of text simplicity, the TC analysis of the CMC sub-corpora was compared with a TC analysis of the same TC issues of a corpus of simplified documents (Simple English Wikipedia (SEW). SEW (Simple English Wikipedia, 2009) is a version of *English Wikipedia* [7], written according to the Basic English controlled language rules (Ogden, 1930). For the analysis, the whole SEW was downloaded, converted to plain text format, and parsed by the *Machinese Syntax* parser (Tapanainen and Järvinen, 1997). The downloaded version of SEW contained 80 067 files, 329 142 sentences and 4 389 599 words.

The results of the text complexity comparison between the whole CMC and SEW for the two metrics of average sentence length (in words) and average word length (in characters), which are considered to be primary metrics reflecting text complexity, are given in Table 2.

The values displayed are the means along with their standard errors.

[7] www.wikipedia.org, last accessed on January 2nd, 2012.

Corpus/ Features	Average sentence length (in words)	Average word length (in letters)
CMC Total	16.211± 0.098	5.462 ± 0.005
Simple English Wikipedia	13.336 ± 0.043	4.764 ± 0.003

Table 2: TC comparison of CMC and SEW.

As can be seen from Table 2, both the average sentence length and the average word length for the CMC are higher than those of the SEW. The results were statistically significant at the 99% confidence level.

The sub-corpora have been compared with Simple English Wikipedia for the following high text complexity features:

- Lexical: average word length (AWL), lexical diversity (LD), average number of word senses (ANWS).
- Syntactic: average sentence length (ASL).
- Discourse: proportion of discourse markers (PDM) and proportion of personal and possessive pronouns (PPPP).

The average word length is an indicator of words that are too long to be process and are an indirect sign of technical vocabulary, which may not be familiar to lay readers. Lexical diversity is an indicator of inconsistent use of terminology, which may confuse the reader, while the average number of word senses indicates ambiguous words. The average sentence length has been deemed to be the best indicator of syntactic complexity (Szmrecsanyi, 2004). The proportion of discourse markers is a formal indicator of well-expressed relationships between separate statements. Finally, the proportion of personal and possessive pronouns (PPPP) is calculated because pronouns cannot be processed by some readers with impaired reading skills, such as patients suffering from aphasia (Canning, 2002). The results are statistically significant at the 99% confidence level and are shown in Table 3.

Corpus/ Features	ASL	AWL	LD	ANWS	PDM	PPPP
1. GP	12.575 ± 0.275	5.114 ± 0.020	0.042 ± 0.001	8.275 ±0.061	0.0015 ±0.000	0.042 ±0.001
2. Spec	16.606 ± 0.160	5.709 ± 0.008	0.017 ±0.000	7.082 ±0.018	0.0012 ±0.000	0.009 ±0.000
3. SC	13.908 ± 0.268	5.222 ± 0.014	0.027 ± 0.001	7.857 ±0.041	0.0021 ±0.000	0.009 ±0.000
4. PMM	17.307 ± 0.131	5.285 ± 0.009	0.025 ± 0.000	7.235 ±0.021	0.0014 ±0.000	0.014 ±0.000
SEW	**13.336 ± 0.043**	**4.764 ± 0.003**	**0.022 ±0.000**	**8.026 ±0.012**	**0.0018 ±0.000**	**0.034 ±0.000**

Table 3: Text complexity comparison of the CMC sub-corpora with Simple English Wikipedia (SEW).

As can be seen, for some of the high text complexity features, the CMC sub-corpora have higher values than Simple English Wikipedia, and exhibit some document-type-specific characteristics. For example, the protocols for Specialists and ProMedMail have much higher average sentence length than Simple English Wikipedia. The instructions for the general population (GP) have much higher lexical diversity (i.e. are characterized by high terminology inconsistency), word ambiguity, and number of pronouns than SEW and the other CMC sub-corpora. The documents which have the highest number of discourse markers are the instructions for pilots, which reflects the fact that they indicate the order of actions very clearly. The results clearly indicate that the existing crisis management documents exhibit a high number of text complexity issues; these hinder human comprehension even in non-emergency situations and thus should be simplified.

4.2. Preliminary linguistic analysis of the crisis management sub-language

The third application of the CMC is the study of the language characteristics of crisis management documents for future tuning of natural language processing applications. The linguistic analysis aimed to test the research hypothesis that the crisis management documents exhibit language specificities which could be attributed to them being written in a specific sub-language. The hypothesis was tested by comparing the crisis management corpus with a random sample of a corpus of standard English (the British National Corpus, or BNC). The size of the BNC sample was 1 401 264 words. The two corpora were compared for proportions of nouns (N), verbs (V), adjectives (Adj) and adverbs (Adv) normalized to the total number of words in each corpus, as well as for average sentence length (ASL), which is both one of the main high text complexity markers and a linguistic feature that varies by genre. Table 4 shows the results of the comparison. The results are again shown with the standard errors of the mean and the differences were again significant with at the 99% confidence interval.

Corpus/ Features	N	V	Adj	Adv	ASL
1. GP	0.360 ±0.003	0.141± 0.002	0.084 ±0.002	0.045 ±0.001	12.575 ±0.275
2. Spec	0.423 ±0.001	0.099 ±0.001	0.099 ±0.001	0.026 ±0.000	16.606 ±0.160
3. SC	0.444 ±0.002	0.099 ±0.001	0.080 ±0.001	0.034 ±0.001	13.908 ±0.268
4. PMM	0.388 ±0.001	0.099 ±0.001	0.084 ±0.001	0.039 ±0.000	17.307 ±0.131
BNC	**0.305** ±0.001	**0.120** ±0.001	**0.086** ±0.001	**0.056** ±0.000	**20.246** ±0.133

Table 4: Sub-language comparison of the CMC sub-corpora with the British National Corpus.

Table 4 shows clear differences between the CMC sub-corpora and the BNC sample, which suggests that crisis management documents are written in a sub-language with different characteristics than general English. Further study is necessary in order to better test this hypothesis (McEnery and Wilson, 1996) and induce the specifics of this sub-language. The results also provide clear motivations for any text simplification (Temnikova, et al., 2012) and NLP applications to be tailored to the CM domain sub-language.

5. Conclusions and Future Work

The current paper has described a novel resource in the Crisis Management domain, namely the first English Crisis Management Corpus, its methods of collection, structure, and composition, as well as possible applications of the resource. More concretely, three tasks have been discussed—one aiming to investigate the text complexity phenomena present in crisis management documents, the second one analyzing the characteristics of the language in which crisis management documents are written, and the third consisting of a corpus-based development of controlled language guidelines for effective simplification of crisis management documents. Other possible uses of the resource have also been described.

Future work includes finding collaborations for building NLP applications for the Public Security field, in order to increase the interest of the Natural Language processing community in this sensitive domain.

6. Acknowledgements

The authors gratefully acknowledge the assistance of Prof. Dr. Galina Maneva with the statistical processing of the corpus analysis data.

7. References

Air Line Pilots Association (1977) Aircraft accident report. Human factors report on Tenerife accident. Engineering and Air Safety, Washington D.C.

Canning, Y. (2002) Syntactic Simplification of Text. Ph.D. Theses, University of Sunderland.

Conway, M. et al. (2009) Using hedges to enhance a disease outbreak report text mining system. In BioNLP ’09: Proceedings of the Workshop on BioNLP, pages 142–143, Morristown, NJ, USA, 2009.

Coppola, D.P. (2007) Introduction to international disaster management, Butterworth-Heinemann.

Corvey et al. (2010) Twitter in Mass Emergency: What NLP Techniques can Contribute. NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media (Los Angeles, California, June 2010), 23–24.

Dale, E. and Chall, J. (1948) A Formula for Predicting Readability. Educational Research Bulletin, pp.11-20, 37-54.

Davis, L. E. et al. (2003) Individual Preparedness and Response to Chemical, Radiological, Nuclear, and Biological Terrorist Attacks. Santa Monica, CA.

Doan, S. et al. (2007). The role of roles in classifying annotated biomedical text. Biological, clinical, and translational language processing, pp. 17-24.

DuBay, W.H. (2004) The Principles of Readability. Costa Mesa, CA: Impact information.

Flesch, R. (1948) A new readability yardstick. Journal of Applied Psychology, 32, pp.221-233.

Harley, T. A. (2008) The Psychology of the Language: from data to theory. Psychology Press.

Ireson, N. (2009) Local Community Situation Awareness During an Emergency. IEEE International Conference on Digital Ecosystems and Technologies.

Johnson, E. et al. (1993). 'PoliceSpeak - Police Communications and Language and the Channel Tunnel - Research Report', PoliceSpeak Publications, Cambridge.

Kienzle, J., Guelfi, N., Mustafiz, S. (2010) Crisis management systems: A case study for aspect-oriented modeling. Transactions on Aspect-Oriented Software Development 7

Kiwan, D. et al. (1999) The effects of text comprehension and performance in examinations. Proceedings of BPS London Conference, December, 1999.

Kosseim et al. (1996). Generating grammatical and lexical anaphora in assembly instructional texts. Trends in natural language generation: An artificial intelligence perspective. Springer.

McEnery, T. and Wilson, A. (1996). Corpus Linguistics. Edinburgh: Edinburgh University Press.

McNamara, D.S., et al. (2010). Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes, 47, 292-330.

Ogden, Ch. K. (1930) Basic English: a general introduction with rules and grammar, London,Kegan Paul, Trench, Trubner.

Ogrizek, M. and Guillery, J-M. (1999) Communicating in crisis. Transaction Publishers.

Regester, M. & Larkin, J. (2005) Risk issues and crisis management. A casebook of best practice. London: Kogan Page.

Renahy, J. et al. (2010) Development and Evaluation of a Controlled Language and of a computerized writing assistant “LiSe” to improve the quality and safety of medical protocols. International Forum on Quality and Safety of Health Care. 20-23 April 2010, Nice, France.

Solheim, T. et al. (1992): The “Scandinavian Star” ferry disaster 1990 - a challenge to forensic odontology. International Journal of Legal Medicine 104: 339-345.

Szmrecsanyi, B. M. (2004) On operationalizing syntactic complexity. In: G. Purnelle, C. Fairon and A. Dister, Editors, Proceedings of the seventh international conference on textual data statistical analysis, Presses universitaires de Louvain. II, Louvain-la-Neuve (2004), pp. 1032–1039.

Steinberger R. et al. (2009). An introduction to the Europe Media Monitor Family of Applications. In: Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop.

Tapanainen P and Järvinen T. (1997) A non-projective dependency parser. In Proceedings of the 5th Conference on Applied Natural Language Processing. Washington D.C.

Temnikova, I., Orasan, C. and Mitkov, R. (2012). CLCM—A linguistic resource for effective simplification of instructions in the crisis management domain and its evaluations.

Unwalla M. (2004) AECMA Simplified English. Journal Communicator, Winter 2004.

Winerman, L. (2009) Crisis Communication. Nature, vol. 457, p 376

Extraction of information on activities of persons suspected of illegal activities from web open sources

Christian Fluhr, Aurélie Rossi, Louise Boucheseche, Fadhela Kerdjoudj
GEOLSemantics
32, rue Brancion, 75015 Paris, France
E-mail: christian.fluhr@geolsemantics.com, aurelie.rossi@geolsemantics.com,
louise.boucheseche@geolsemantics.com, ker.fadhela@gmail.com

Abstract

This work is part of the French funded SAIMSI project (Suivi Adaptatif Interlingue et Multisource des Informations). The aim of the project is to follow activities of persons suspected of illegal actions like terrorism, drug traffic or money laundering. The paper focuses on the information extraction in particular. This extraction is done in French, English, Arabic and Chinese. The information extraction is based on a deep morphosyntactic analysis. Recognition of single words, idiomatic expressions, compounds is performed and named entities are identified and categorized. Dependency relations are built, passive/active forms, negation anaphora, verb tenses are processed. Information extraction is application-independent and uses extraction rules. At this level some named entity categories can be reconsidered. This extraction is based on a large security ontology. The paper details the problems of the consolidation of the extracted knowledge at the document level. The future evaluation on WEPS-3 data is presented.

Keywords: Information extraction, illegal activities, RDF, web open source information

1. Context

This work is part of the SAIMSI [2] project funded by the French Research Agency. The aim of the project is to follow activities of persons suspected of illegal activities (terrorism, drug, money laundering, etc). If possible, the activities must be located in time and place (geochronolocalization).

The processed information is taken from open sources on the Internet (news sites, social networks, specialized sites) in 4 languages: French, English, Arabic and Chinese (Mandarin).

The extracted information is structured and coded into RDF in English whatever the source language, according to security ontology. This means that information coming from various languages can be merged into a knowledge base.

As the project concerns international activities, a special attention has been paid to the different spelling of person names originally coded in different character sets. This gives a better recall but increases the problem of homonymy resolution.

With the result of knowledge extraction, two databases are built.

A knowledge base that stores the triples allows reasoning to infer new relations and controls the consistency of the knowledge. The results of queries are presented in biographic sheets, geographic maps, timelines and graphs of relations between persons and/or organizations.

A cross-language text database is also built. It is used to control the origin of the extracted knowledge and gives the possibility for the user to interrogate on themes that have not been structured into the knowledge base.

This paper focuses on the extraction of knowledge based on the result of a deep general-purpose morphosyntactic analysis and an application-oriented knowledge extraction using rules. The incomplete information obtained from sentences is discussed. The complementation of the extracted knowledge by a local reasoning on the full document is presented.

This paper does not describe the introduction of new document knowledge into the knowledge base that contains the instances and properties already introduced from previous documents and especially the processing of homonyms. It is an ongoing work.

2. Related works

Developments made by the Joint Research Center (JRC) of the European Commission in Ispra [5] are the closest work to ours. This work (EMM) consists in gathering of news in various languages, processing of name variants in different character sets, construction of graphs of personal links and event extraction. The common work with FRONTEX [1] and the University of Helsinki is security-oriented. The aim of this project is to extract, in 7 languages, border security oriented events. Two approaches have been tested.

The first one (NEXUS) by JRC begins by a clustering that gathers texts arriving in a 10 minute window into groups relating to the same event. After a morphological analysis extraction rules are performed on the beginning of each text. They consider that the main information is in the beginning of texts and because of redundancy, lost information in a document can be extracted in the beginning of another.

The other approach is provided by the University of Helsinki (PULS) [6] and performs a morphosyntactic parsing of the full document including resolution of anaphora. This approach is close to ours but the problem discussed in this paper about consolidation of knowledge at the document level seems to be not taken into account.

A large amount of police and intelligence services use the tools from I2 (I2 base, Analyst's Notebook) [3]. Extraction and categorization of named entities done by TEMIS have been introduced into TextChart AutoMark helping users to fill the Base.

3. The Ontology

The knowledge to be extracted is described in a security ontology .

For each person, a list of attributes has to be extracted: given names, surnames, middle names, nicknames, birth dates, birth places, diplomas, personal addresses, fixed telephone numbers, mobile phone numbers, e-mail addresses, Web site, internet domain, etc.

Actions to be extracted are: travelling from a point to another, contacts between persons and/or organizations, payment or purchase, construction of objects, emission of a message (discourse, book, mail, interview, etc.), events of the life (birth, marriage, divorce, death, funeral), family relations (brother, sister, son, daughter, grandson, etc.), interaction with police and justice (control by police, arrest, conviction by a court, release from prison, etc.), link and role with an organization, teaching organization attended.

This ontology has been built after discussions with users. The main applications that have been considered are terrorism and malware production and use (especially phishing).

4. Deep morphosyntactic analysis

In order to minimize the effort to write extraction rules, we have chosen to build them on the result of a deep morphosyntactic analysis which is domain independent. That means that our extraction rules are built by linguists. There are no extraction rules built on the surface level that can be learned from a tagged corpus. This minimizes the number of rules, because they are applied to a deep representation that is independent of the surface level.

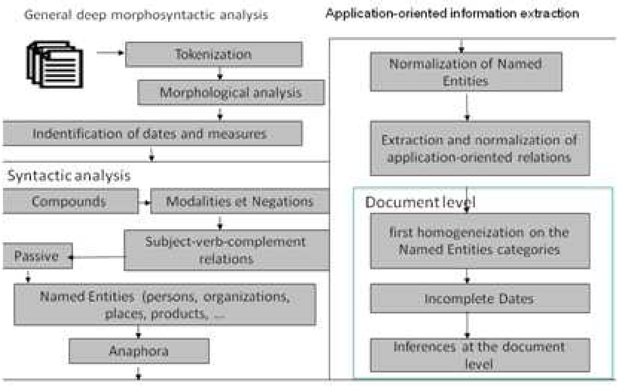

General deep morphosyntactic analysis
Application-oriented information extraction
Tokenization
Morphological analysis
Indentification of dates and measures
Syntactic analysis
Compounds
Modalities et Negations
Subject-verb-complement relations
Passive
Named Entities (persons, organizations, places, products, ...
Anaphora
Normalization of Named Entities
Extraction and normalization of application-oriented relations
Document level
first homogeneization on the Named Entities categories
Incomplete Dates
Inferences at the document level

Figure 1: schema of the linguistic processing and information extraction

The morphosyntactic analysis is developed using a linguistic-oriented programming language that is executed on a weighted finite state automata engine.

The morphological level recognizes words even for Mandarin language which has no space character between words. Syntactic and some semantic information is associated using a dictionary of single word forms and a dictionary of idiomatic expressions.

Some named entities are recognized and normalized at this level such as dates, phone numbers, e-mail addresses, etc.

A part of speech tagging then disambiguates grammatical categories and helps to choose the lemma and more generally the normalization of each word. Normalization is a common representation of different lemmas that are full synonyms or spelling variants. It is the case for UK and US spellings like “colour” and “color”.

The syntactic parsing recognizes dependency relations inside noun and verb phrases as well as subject-verb-object relations, and gives a common representation of active and passive forms by producing agent-action-object relations. Verb tenses including compound tenses are recognized, negation is also recognized along with modalities. Pronouns and possessive adjectives are processed. This last point is very important because without pronoun recognition a lot of knowledge cannot be extracted.

Pronouns are mainly processed at the paragraph level, but for particular type of documents like biographies it is necessary to process pronouns at the document level.

During this step, the rest of named entities are recognized and typed using both recognition rules and lists of known entities coming from databases like Geonames or DBpedia.

5. Extraction

The extraction rules are written in the same language as the one used for syntactic analysis. They are triggered using words representing different actions or properties described in the ontology. The rule verifies that it is really the supposed action and then recognizes and links different role players (agent, object, place, date, instrument, manner, etc).

Verification is necessary because some triggers can be ambiguous. For example, in French “se rendre” can be “to go” or “to surrender” that represent different actions in the ontology. Of course, if the agent is in both cases a person, the other role players have different types.

Other kinds of semantic ambiguities are processed using these rules, like “Father” which can be an ecclesiastical title or a family relation.

Example of extraction:

The result is RDF-coded but in order to minimize the place in this paper we will give the natural language reconstruction of the RDF. This natural language reconstruction of an English oriented RDF can be done in any language even if no morphosyntactic parsing is available for this language. It’s a way to have a cross-language summary of the activities interesting the users.

***Original sentence** : “Basam Ayachi a organisé le mariage de Malika el Aroud avec Abd el Satar Dahmane qui a tué le Commandant Massoud le 9 septembre 2001.“*

***Result** :*

Union organised by Basam Ayachi between Malika el Aroud and Abd el Satar Dahmane.

ViolentAct author Abd el Satar Dahmane, victim Commandant Massoud date 20010909

Death Commandant Massoud date 20010909 type: murder

The last information is not really in the sentence but is inferred from the type of violent act.

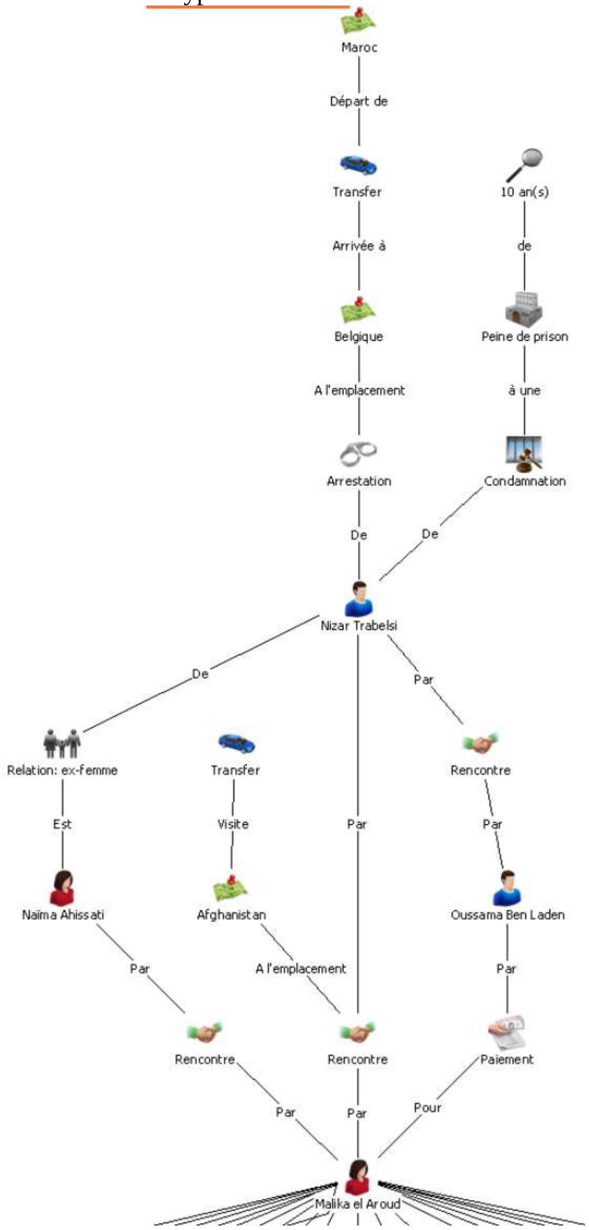


Figure 2: Visualisation of extracted information using Analyst Notebook of I2.

6. Limitation of the knowledge extraction from sentences

As you can see in the previous example, information which should be important is lacking in the sentence like the date and place of the marriage and the place of the Massoud assassination. It lacks also the given name (Ahmed Shah) of Massoud. All this information can often be obtained from the same document.

In fact, the extraction must be consolidated at the document level. Documents have a semantic coherence. The discourse is constructed to prevent the reader from having confusions, but at the same time some information is not repeated because the reader can infer it.

That is the reason why it appears compulsory to use the order of succession of sentences in the documents and also the information about the publication date of the documents given by the metadata, because they are used by the reader to understand the text.

6.1 Identification of persons and places

Persons and places are cited in different sentences with sometimes different forms. If the author wants to give information about two different people having the same surname, he will prevent confusion by giving distinctive information (given name, Jr/Sr, title). On the contrary, compatible names can be gathered using the same identification number.

6.2 Relative dates

Some relative dates depend on the publication date. In this case the tense of the verb is important. “On Monday, Jack went to London”/ “On Monday Jack will go to London”, in these examples, Monday refers to different dates.

As dates can be fuzzy in natural language, they are represented in RDF as slice of time representing the incertitude about the event date. A date is represented by two dates that are equal in case of full date. For example, “2007” is represented by datebeg=20070101 and dateend=20071231.

Some dates are relative to national or religious dates. For example, “next Easter” needs to get the date for this year. This kind of event has dates changing each year like the Ramadan period.

Other relative dates depend on a date given previously or after in the text. Example: Massoud was murdered two days before the 9/11/2001.

6.3 Itineraries

Some texts can contain the description of an itinerary. Each sentence gives a part of the itinerary, but each sentence extraction gives partial information.

Example: “John went to Istanbul on August 25 2004. Two days later, he arrived in Ankara. “

6.4 Some dates and/or places can be linked with following sentences

“John went to Lisbon on Monday. He met Jack. “

The extracted knowledge from the first sentence is quite complete. It lacks only the departure place which can be John’s residence place. But in the second sentence, the extraction gives only a meeting between John and Jack but no place and no date.

If no language construction prevents from understanding that it is the same place and the same date, the geochronolocalization information from the first sentence must be attributed to the second one.

This kind of reasoning cannot be done without a study which is an ongoing process in our company.

In what cases this inference can be done? With what succession of actions?

An examination of this problem on a corpus is done to find the situations where inference must be applied.

6.5 Processing of well known persons

For some well known people, anaphoras can be done using known particularities like the title of the person which is not linked in the document with the noun.

For example: “Sarkozy went to Marseille. After that, the president went to Nice.“

The fact that Sarkozy is the president is not mentioned in the document because it is well known during a time period. To solve this difficulty, databases on well known persons or organizations are used, like DBpedia. This action is under development. DBpedia files which have some errors are corrected manually before being used to tackle this problem of anaphora.

7. Evaluation

7.1 Inner evaluation

GEOLSemantics has developed its own evaluation system. This evaluation system is also used as non regression test between successive versions of the system. Texts are processed by the last successive versions. The RDF result is splitted into triples. Each triple is manually controlled if it has not been accepted or rejected by a previous version. Triples which are lacking are added by the evaluator. They are not added in RDF which is too heavy but are just reported by a message giving the waited information. When a new version gives the right triple, the text one is suppressed.

With this tool, progression in output quality can be continuously estimated.

7.2 Evaluation campaign

The closer evaluation campaign is WEPS. There are no more campaigns organized but WEPS-3 data is available to test our system. Unfortunately, WEPS is only about attributes on persons without any attempt to extract actions.

We intend to test our system against WEPS-3 Data [4] by the end of this year. Our approach will be the extraction of all the relations we are able to extract and their use to perform a better clustering. Some extracted attributes are incompatible, for example birth place or date. Persons having incompatible attributes are considered as different persons.

In the case where there is no incompatibility, the proximity between persons will be calculated using the other attributes and the vocabulary contained in each documents. Vocabularies concerning a singer, a football player or a medicine doctor are strongly different.

The similarity matrix necessary for a clustering can benefit from the combination of the incompatibilities and the vocabulary contained in the texts.

The examination of the WEPS-2 data shows that there are different types of documents that should be processed in different ways.

For example, some documents are lists of homonyms. For each homonym, there is a list of id discriminative information like birth date, birth place, telephone number, address, etc. These documents must be splitted into subdocuments for each different homonym.

8. Conclusion

This work is an ongoing process. Information extraction in the French language is quite finished. English will be fully developed by August, Arabic and Chinese by the end of the year. This information extraction is included into the full SAIMSI process based on the WEBLAB platform from Cassidian.

Application for semiautomatic filling of Ibase (I2) is possible.

9. Acknowledgements

This work has been done by the support of the Agence Nationale de la Recherche, Project SAIMSI (ANR-9-CSOSG-08-01).

10. References

[1] Atkinson M., Piskorski J., Frontex Real-time News Event Extraction framework, *conference KDD’11*, August 21-24, 2011, San Diego, California, USA

[2] Fluhr C., SAIMSI, Suivi Adaptatif Interlingue et Multisource des informations, *workshop WISG’12*, 24-25 janvier 2012, Troyes, France

[3] I2 Limited, TextChart AutoMark 8, 2010

[4] Martin N., Khelif K., Focussed crawling using name disambiguation on search engine results, *International Symposium on Open Source Intelligence & Web Mining*, 12-14 September 2011, Athens, Greece

[5] Tanev H., Piskorski J., Atkinson M., Real-time News Event Extraction for Global Crisis Monitoring, *conference NLDB’08, processing of the 13th International Conference on Natural Language and Information Systems*, Springer Verlag Berlin, 2008

[6] Yangarber R., Jokipii L., Rauramo A. and Huttunen S., Extracting information about Outbreaks of infectious Epidemics, In Proceedings of the HLT-EMNLP 2005, Vancouver, Canada, 2005.

Language Processing and Linguistic Data in the CAPER Project

Carlo Aliprandi, Tomas By, Sérgio Paulo

Synthema
Via Malasoma, 24
56121 Pisa, Italy

Vicomtech IK4
Mikeletegi Pasealekua, 57
20009 Donostia / San Sebastián, Spain

Voice Interaction
Rua Alves Redol, 9
1000-029 Lisboa, Portugal

E-mail: carlo.aliprandi@synthema.it, tby@vicomtech.es, spaulo@voiceinteraction.pt

Abstract

Much information of potential relevance to police investigations of organised crime is available in public sources without being recognised and used. Barriers to the simple and efficient exploitation of this information include that not everything is easily searchable, and may be written in a language other than that of the investigator. To help overcome these problems, the CAPER project aims to create an integrated platform for acquisition, processing, and analysis of information in multiple languages, and also link this to legacy police IT systems. Full Natural Language Processing pipelines for multiple languages and media are used to map persons and organisations to actions and events, and Multi-lingual lexicons and gazetteers allow cross-lingual search in the indexed data. Domain-specific lexicons contain words and slang expressions with special senses in the context of organised crime. The system supports multilingual analysis of unstructured and audiovisual contents, based on text mining for fourteen languages, and uses language-neutral interfaces, so that addition of further languages will not require any modification of existing components.

Keywords: Open Source Intelligence, organised crime, cyber crime.

1. Introduction

The EU FP7 CAPER project (Collaborative information Acquisition, Processing, Exploitation and Reporting for the prevention of organised crime) uses state-of-the-art, multi-lingual Natural Language Processing techniques to support analysis and sharing of information obtained through search and monitoring of Web data (Aliprandi & Marchetti, 2011).

CAPER will allow Law Enforcement Agencies (LEAs) to share informational, investigative and experiential knowledge. A common software architecture for all linguistic processing components allows efficient combination of language-specific and language-independent modules. Cross-lingual search and query expansion is supported by multilingual lexicons and gazetteers. The project includes full support for English, Spanish, Catalan, Italian, and Portuguese, and partial support for French, German, Romanian, Russian, Basque, Hebrew, Arabic, Chinese, and Japanese.

2. Open Source Intelligence and Security

Internet invaded our lives to such an extent that words like e-commerce, e-learning or e-government became familiar to many of us. Moreover, the widespread use of internet resulted in the emergence of highly interconnected societies. From the LEAs point of view, while such societies pose new challenges, they also provide new collaboration opportunities. That is, on the one hand, organised crime can use information technology systems to communicate, work or expand their influence to anywhere in the planet, but current tools for fighting such organisations have shown their limits and reflect the need for developing a scalable tool to track them more efficiently. On the other hand, for languages addressed in the project, internet became a massive repository of both written and spoken data (audio and video files have enjoyed increasing success as a means of information dissemination over the last few years) and, thus, comes up as a priceless source of material to be searched while attempting to detect potentially criminal activities.

CAPER's objective is to build a common collaborative and information sharing platform for the detection and prevention of organised crime in which the Internet is used (e.g. sale of counterfeit or stolen goods, cyber crime) and which exploits Open Source Intelligence (OSINT).

The benefits of using OSINT for Intelligence Agencies (IA) are well known from previous experiments. LEAs, alike IAs in the past, are increasingly more reliant on information and communication technologies and affected by a society shaped by mass media, but more and more by the Internet and social media. This challenge can also be seen as an opportunity for LEAs. The richness and quantity of information available from open sources, if properly processed, can in itself provide valuable intelligence and help draw inference from existing closed source intelligence. The CAPER project is aimed at giving evidence that OSINT can help LEAs better understand the information they have available: the talk will detail the CAPER platform elements, giving particular evidence from a user-oriented perspective to the exploitation of Open Data Sources and the integration of mass media, closed information sources (LEA internal systems) and Social Web data sources, like Linked Data, Wikipedia, Geonames, or Yago.

A typical use case for the CAPER system is when an investigator uses the Internet as a reference source to extract structured and unstructured information of relevance to an ongoing criminal case. The investigator can establish mechanisms to automate the daily work; i.e. periodic access to predetermined sites, downloading predefined content, manage users and passwords, set alerts, identify individuals with multiple identities in the net etc. The system supports cross-lingual search and indexing, and will be linked to third-party machine-translation services for on-demand translation of texts and documents.

3. The CAPER system

The project is not focused on developing new technology, but on the fusion and real validation of existing state of the art to solve current bottlenecks faced by LEAs. So the design methodology will be based on an iterative development lifecycle, with several integration steps that will be carried out.

The 6 technology pillars of the CAPER platform are:

- Open and Closed Data Sources: TV, Radio, and Information in closed legacy systems are the data sources to be mined and evaluated by CAPER, in addition to Open Internet data sources and Semantic Web data collections (Linked Data, for example, like Geonames, DBpedia or Yago).
- Data Acquisition: Depending on the information source type, different acquisition patterns will be applied to ensure acquired information is the richest possible and has a suitable format for analysis.
- Information Analysis: Each analysis module is geared towards a specific content type, i.e. Text, Image, Video, Audio and Speech or Biometric data. These modules interact with a ‘Semantic mash-up’ component, to interlink Semantic Web data.
- Information and Reference Repositories: source data and mined information will be stored in these repositories, separated by content type. Repositories will also store the reference images, text, keywords, biometric data etc. of interest to the LEAs.
- Interoperability and Management Application: This is the end users’ workbench, the main Human Computer Interface. It will allow LEAs to collaborative create and configure their monitoring requests and analysis petitions. Through this HCI, Law Enforcement Officers will be able to create and configure their monitoring requests and analysis petitions. It will allow a structured collaboration between LEAs, who will also be able to configure their own internal closed information sources and control how and to whom the data is shared.
- Visual Analytics (VA) and Data Mining (DM): VA and DM will provide the intelligence necessary to support the output of the system. They will allow LEAs to effectively mine processed data both from Closed and Open Sources, and to further relate it to Semantic Web sources when required.

4. Natural Language Processing in CAPER

The CAPER system will be designed from a linguistically neutral point of view. This design methodology will allow linguistic analysis and speech recognition components for any language to be added in the future. LEA users will provide reference images, keywords, biometric data, and define concepts to be used in information acquisition.

The design methodology will leverage standardization of data and interchange of tools: once a data format is accepted as a standard, tools can be adapted and shared with little effort. CAPER will also standardize the processing of language sensitive information by adopting and extending the Knowledge Annotation Format (Bosma, 2009), a multi-layered XML format for linguistic and semantic annotation of unstructured documents that has been shown to be suitable for the purpose of information processing. CAPER aims at extending data representation standards to also cope with multimedia and structured data, to integrate Social Web and Semantic Web content.

Processing components vary depending on availability of state-of-the-art systems for each language. For the first prototype, the CAPER system will use OpenNLP for English, Freeling for Spanish and Catalan, Synthema SG for Italian, and XIP for Portuguese.

Freeling (Padró et al., 2010) includes a chart-based shallow parser and a rule-based dependency parser. The dependency parser works in three stages: first parsing rules are used to complete the shallow parsing produced by the chart into a complete parsing tree. These rules are applied to a pair of adjacent chunks. At each step, the selected pair is fused in a single chunk. The process stops when only one chunk remains. The next step is an automatic conversion of the complete parse tree to a dependency tree. Since the parsing grammar encodes information about the head of each rule, the conversion is straightforward. The last step is the labeling. Each edge in the dependency tree is labeled with a syntactic function, using the labeling rules.

Synthma SG (Aliprandi et al., 2008; Neri et al., 2011) is a multilingual rule-based parser, performing in one single step multiple NLP tasks that include: document and sentence segmentation, word tokenization, Part-of-Speech tagging, lemmatization, Chunking, Named Entity Recognition, Dependency Parsing, co-reference resolution and Semantic Role Labeling. The parser is the core technology for a wide range of applications, including Text Mining and Machine Translation. Syn ESG is intended to identify relevant knowledge from the whole raw text, by detecting semantic relations and concepts in texts. Concept extraction and text mining are applied through a pipeline of linguistic and semantic processors that share as a common ground McCord's theory of Slot Grammar. A slot is a placeholder for the different parts of a sentence associated with a word. A word may have several slots associated with it, forming a slot frame for the word. In order to identify the most relevant terms in a sentence, the system analyzes it and, for each word, the Slot Grammar parser draws on the word's slot frames to cycle through the possible sentence constructions. Using a series of word relationship tests to establish the context, the system tries to assign the context-appropriate meaning (sense) to each word, determining the meaning of the sentence. Each slot structure can be partially or fully instantiated and it can be filled with representations from one or more statements to incrementally build the meaning of a statement. This includes most of the treatment of coordination, which uses a method of ‘factoring out’ unfilled slots from elliptical coordinated phrases. Syn ESG parser - a bottom-up chart parser - employs a parse evaluation scheme used for pruning away unlikely analyses during parsing as well as for ranking final analyses. By including semantic information directly in the dependency grammar structures, the system relies on the lexical semantic information combined with functional relations. Beside Named Entities, locations, time-points, etc, Syn ESG detects relevant information like noun phrases which comply with a set of pre-defined morphological patterns and whose information exceeds a threshold of significance. The detected terms are then extracted, reduced to their Part Of Speech and Functional tagged base forms.

XIP (Paulo et al., 2008) consists of a cascaded finite state parser using a set of ordered grammars, providing a formalism that smoothly integrates a number of description mechanisms for shallow and deep robust parsing. As the input data flows sequentially through the grammar pipeline, it is either enriched or left unchanged. Moreover, each rule of a given grammar can refer to representations produced by processing the preceding grammars. The Portuguese-specific grammars were jointly developed by the Spoken Language Systems Lab and Xerox the system is available through commercial license.

The expected final output of the NLP pipeline(s) consists of entities (names, nouns, pronouns) and dependency links between these and the verbs. This is not necessarily a complete description of the syntactic or semantic structure of the sentence, but a simple representation that can be computed with reasonable robustness. The main purpose is to offer cross-lingual search on names and concepts. Later in the project, word sense disambiguation and semantic role labelling will be integrated, with a view to constructing event representations (Aliprandi et al., 2011; Hagège et al., 2009).

5. Multi-lingual indexing and search

To support access of content in one language using search terms from another language, the CAPER system will need mappings between languages. The Multilingual Central Repository (Atserias et al., 2004) provides a mapping between content words (nouns, verbs, adjectives, adverbs) in English, Spanish, Catalan and Basque. Each word is linked to one or more concepts (synsets) and synonymous words, within one language and between languages, share the same concepts. The EuroWordNet database, available for a fee, contains similar mappings for Italian, Spanish, German, and French, and there is also a separate French wordnet project, freely available. For names, or named entities, the JRC-names database (Steinberger et al., 2011) contains a mapping from text strings to numerical identifiers for the entities, together with language information. While most names are language-independent, in some cases a name in one language can also be a content word in another language, and in some cases the same name is spelled slightly differently in different languages.

The Multilingual Central Repository (MCR) is a collection of WordNet lexicons and several ontologies, all linked to the same concepts (synsets). An `Inter-Lingual-Index´ in the MCR connects words in one language to equivalent translations in any of the other languages.

JRC-Names is a highly multilingual named entity resource for person and organisation names. It consists of large lists of names and their many spelling variants (up to hundreds for a single person), including across scripts (Latin, Greek, Arabic, Cyrillic, Japanese, Chinese, etc.). The software is implemented in Java and analyses UTF8-encoded text files to find known entity mentions, returning the name variant found, the preferred display name for that entity, the unique name identifier for that name, the position of the entity name in the text, and its length in characters.

6. Security-specific Language Resources

The domain-specific linguistic resources used in CAPER include extensions to general lexicons and ontologies, for normal words that have special senses in certain contexts, such as the drugs trade, as well as slang expressions not normally used in the language. There will also be interpretation/conversion modules for the special jargon and spelling conventions used in online chat rooms and similar environments.

The Multilingual Lexical Database in CAPER will also contain named entities (NEs) from various sources. Most importantly, there will be provision for LEA’s to enter their own collections of names of interest. For toponyms, we plan to compile a multilingual gazetteer of place names from various sources, and, if needed, provide access to the Google maps API. Furthermore, Wikipedia contains extensive lists of NEs, but some manual work will be needed to classify these, and it is not clear if this source will provide much that is not already in one of the other collections already mentioned.

Additionally, there will, in CAPER, be provisions for LEAs to add their own terms and concepts, connected to the MCR, and share them with each other confidentially.

7. Summary

CAPER will allow investigators to seamlessly search and integrate information in foreign languages (among those supported by the system), and thus more effectively assess threats and interpret intelligence. There is much information readily available on the Internet and in the Mass Media, of relevance to investigations but that is missed because investigators are limited to their own language, and may not be aware of related investigations in other countries. CAPER can thus help investigators better understand the information they have available.

8. References

Aliprandi C., & Marchetti, A. (2011). Introducing CAPER, a Collaborative Platform for Open and Closed Information Acquisition, Processing and Linking. Proceedings of HCI International. CCIS 173. Heidelberg, Germany: Springer, pp. 481-485.

Aliprandi, C., Neri, F., Lotti, L., & Sanna. G. (2008). Online Police Station, a cutting edge service against cybercrime. In *Data Mining IX: Data Mining, Protection, Detection and other Security Technologies*. Southhampton: WIT Press, pp. 325-334.

Aliprandi, C., Ronzano, F., Marchetti, A., Tesconi, & M., Minutoli, S. (2011). Extracting events from Wikipedia as RDF triples linked to widespread Semantic Web Datasets. Proceedings of HCI International. LNCS 6778. Heidelberg, Germany: Springer, pp. 90-99.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., Vossen, P. (2004). The MEANING Multilingual Central Repository. Proceedings of GWNC. Brno, Czech Republic: Masaryk University, pp. 23-30.

Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., & Apiprandi, C. (2009) KAF: a generic semantic annotation format. Proceedings of the 5th International Conference on Generative Approaches to the Lexicon.

Hagège, C., Baptista, J., & Mamede, N. (2009). Portuguese Temporal Expressions Recognition: from TE characterization to an effective TER module implementation. In The 7th Brazilian Symposium in Information and Human Language Technology.

Neri F., Aliprandi C., & Camillo F. (2011) Mining the Web to monitor the Political Consensus. In *Counterterrorism and Open Source Intelligence.* LNSN 2. Heidelberg: Springer, pp. 391-412.

Padró, L., Collado, M., Reese, S., Lloberes, M. & Castellón, I. (2010). FreeLing 2.1: Five Years of Open-Source Language Processing Tools. Proceedings of Language Resources and Evaluation Conference.

Paulo, S., Oliveira, L. C., Mendes, C., Figueira, L., Cassaca, R., Viana, C., & Moniz, H. (2008). DIXI –A Generic Text-to-Speech System for European Portuguese. In PROPOR 2008. Heidelberg, Germany: Springer, pp. 91-100.

Steinberger, R., Pouliquen, B., Kabadjov, M., Belyaeva, J., & van der Goot, E. (2011). JRC-Names: A freely available, highly multilingual named entity resource. Proceedings of RANLP. Shoumen, Bulgaria: Incoma Ltd., pp. 104-110.

Tesconi, M., Ronzano, F., Minutoli, S., Aliprandi, C., & Marchetti, A. (2010). KAFnotator: a multilingual semantic text annotation tool. Proceedings of the 5th Workshop on Interoperable Semantic Annotation.

Detection of Child Sexual Abuse Media on P2P Networks: Normalization and Classification of Associated Filenames

Alexander Panchenko, Richard Beaufort, Cédrick Fairon
Centre for Natural Language Processing (CENTAL) – Université catholique de Louvain
Place Blaise Pascal 1 – 1348 Louvain-la-Neuve – Belgium
E-mail: {alexander.panchenko, richard.beaufort, cedrick.fairon}@uclouvain.be

Abstract

The goal of the iCOP project is to build a system detecting the originators of pedophile content on P2P networks such as BitTorrent, eDonkey, or Kad. This paper outlines the key functions of the language processing in the iCOP system. Next, we describe the architecture of the language analysis module and its key components – filename classifier, term extractor, and filename normalizer. The language resources used in each component are discussed. The paper is also presenting the first experiments with the module on the standard porn data (used in the preliminary tests as a substitute of child pornography data). Our results show that the module is able to separate titles of the pornographic galleries and videos from the titles of encyclopaedia articles with accuracy up to 97%. Finally, we discuss the directions for the future research and developments of the iCOP language analysis module.

Keywords: text classification, text normalization, P2P networks, CSA, porn filters

1. Introduction

The goal of the iCOP project[1] is to develop a novel forensics software toolkit to help law enforcement agencies across the EU identify new or previously unknown child abuse media and its originators on peer-to-peer (P2P) networks. Until now, the only way to identify such media is through manual analysis by law enforcement personnel. However, such a manual approach is difficult or impossible given the large number of files that need to be reviewed individually. The limited resources that law enforcement agencies possess make it impractical for them to examine the thousands of new files that may appear on P2P networks every day.

The key output of the project – the iCOP software toolkit – will be used by law enforcement to help detect, filter, and prioritize new instances of child abuse media on P2P networks. iCOP system is operates alongside existing P2P monitoring tools like PeerPrecision. Using trace data from these monitors as input, iCOP identifies candidate suspect media that contain Child Sexual Abuse (CSA) based on a combination of advanced pattern recognition techniques. The language analysis tool employs potential CSA media in P2P networks based on a combination of several sources of evidence: the modelling of offender file sharing behaviour; the analysis of file sharing patterns and query patterns; language processing of file names and queries. Candidate media discovered by language analysis are fed to a content-based media analysis tool, which prioritizes and filters material further.

The core components of the iCOP system are targeted at an automatic identification of new CSA content and its distributors, using evidence in the form of textual queries and filenames in P2P networks, of user's file sharing behaviour, and of the image and video content itself. It is of particular importance to search for “new” material as these previously unknown files are usually introduced on the network by people who one believed to be close to the victim or the abuse. These files have also a greater chance to be related to on-going situation of abuse.

Our contribution to the iCOP is focused on the language technology. In particular, we are working on machine learning techniques that recognize the pedophile queries or filenames in the data of a P2P system. The development of these components follows an iterative optimization process in which the underlying statistical models and data representations are modified and system performance is validated on a test dataset. In iCOP, such a data-driven improvement is not straightforward, as number of target texts (filenames and queries) is very limited and may even be illegal to access in some cases. Furthermore, crawling additional pedophile texts directly from the network is illegal. Therefore, the construction of appropriate training (language) resources require special attention. Evaluation and optimization strategies need to be adapted to work around the problem of accessing the data.

In the following section, we outline the architecture of the language processing module of the iCOP system. We also discuss the assessment strategies with respect to the file classification, term extraction, and filename normalization components of the module. Next, we present our first experiments with the module. Results of the classification of the “standard” porn texts are described. We conclude with directions for the future research and developments.

[1] http://scc-sentinel.lancs.ac.uk/icop/

2. Architecture of the System

Figure 1 presents the architecture of the language analysis module and outlines how it will be integrated with the iCOP toolkit[2]. The core goal of the this module is to complement the media and behaviour analysis tools of iCOP with an analysis of filenames and other text content associated to a candidate child sexual abuse media (i.e. metadata enclosed with videos and pictures). Other modules of iCOP will identify potential CSA images and videos. These files are used as input for the language analysis module. First, the module checks if the media is a known CSA file. If not, features are extracted from the input file name and its text description. They are used in a statistical model which decides if the file contains CSA materials. Secondly, the module extracts key terms describing the CSA media. These terms are used to analyse and track the evolution of the offender vocabulary over time.

The iCOP toolkit is going to use the language analysis module to analyse both filenames and queries. Our module processes these two kinds of texts exactly in the same way. In the following, we will refer to language analysis as filename processing. In the next sections, we will outline methods used to build the file classification, the term extraction and filename normalization components of the language analysis module.

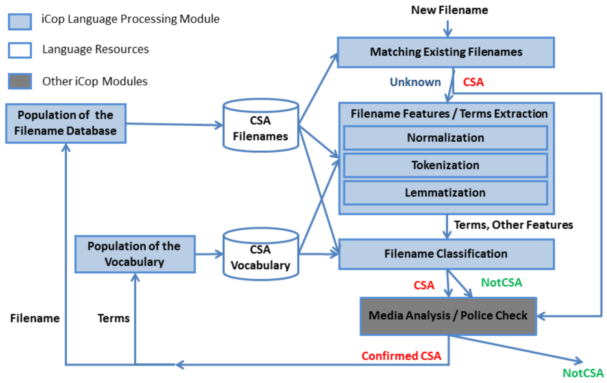


Figure 1: Architecture of the language analysis module of the iCOP forensics toolkit.

2.1 File Classification by Textual Description

The goal of the file classification component is to decide if an input file contains some CSA material using solely the textual descriptions of the file. This kind of classification is challenging for several reasons. First, filenames may be meaningless, such as "0012664321.mpg". Second, file text descriptions and metadata may be short or absent. Finally, text descriptions often use highly non-standard language (abridgments, abbreviations, spelling errors, technical terms, etc). It is also challenging to separate CSA from regular porn as the two categories use a lot of common terms.

In order to deal with these and other issues we have developed a text classification system. The purpose of this system – given a file name and its metadata as input – is to check if the file is an already known CSA media stored in a database of filenames. If the file is unknown, features are extracted from the text describing the file. To do so, the text is segmented. Next, text normalization (Beufort et al., 2010) is used to represent content of a file from scarce, noisy, and misspelled text descriptions. Then a statistical machine learning classifier (a Support Vector Machine or a Regularized Logistic Regression) is used to separate regular files from those containing some abuse content. We rely on the LIBSVM (Chang and Lin, 2011) and LIBLINEAR (Fan et al., 2008) classification software for two reasons. First, because of their commercial-friendly open source licenses. Second, because the libraries efficiently implement the state-of-the-art classification algorithms.

The classifier is trained on a database of file names containing CSA material. The file classification module is used to recognize newly published files in a P2P network which contain abusing content. The list of these candidate files is transferred to the Media Analysis component, where the content of each file is further analyzed (Ulges and Stahl, 2011).

The file classifier is trained and tested on three sets of filenames: regular filenames, regular pornographic filenames, and filenames containing CSA material. We constructed the two first datasets ourselves from the files openly and legally available on the Internet. The third dataset should be provided by our collaborators from law enforcement. Each dataset is a list of entries composed of tuples <Class, Filename>. Here, the Filename is the original name of the file (with the tags if any); Class is either "positive" (porn) or "negative" (non-porn). Following the standard machine learning methodology, these datasets (divided into disjoint training and validation sets) are used to assess performance of the classifiers. Cross-validation is used to estimate performance of the filename classification. In particular, the following metrics are calculated: accuracy, balanced classification rate, mean squared error (MSE), root-mean-square error (RMSE). We present the first results of filename classification in Section 3.

2.2 Term Extraction

For each detected child abuse media, our module returns a set of normalized terms describing it. These terms let monitors seamlessly analyze and track topics on P2P networks. Basically, the extracted terms, are normalized lexical units obtained at the feature extraction stage. Words which were never used before for describing CSA content are also recognized. Detection of the new terms aims to help identify personal names, and locations related to recent victims and abuse cases. In order to detect new terms, we rely on a database of CSA terms. A term is considered new if and only if it appears in a CSA filename (alongside known CSA terms) but is not present in the CSA vocabulary.

The database of CSA language is constructed during the training phase. It is based on specialized hand-crafted terminologies provided by law enforcement agencies. This initial lexicon is enriched with terms extracted from

[2] http://cental.fltr.ucl.ac.be/projects/icop/

the filenames containing child abuse media. To do so, we run the segmentation and normalization tools used on the feature extraction stage on these files and add the most frequent terms to the CSA vocabulary. The CSA database is updated once new files containing illegal material are provided.

The evaluation of the term extraction relies on the users' feedback. Once the module is integrated into the iCOP toolkit, the users can interact with it. Thus, this evaluation will be conducted at the final stages of the project. Generally, satisfaction of users will be an evaluation measure for this module.

2.3 Filename Normalization

The text normalization component module transforms input filenames into normalized versions. Text normalization plays two important roles in the system. As we have already said, filenames and their text descriptions contain highly non-standard language patterns, such as abridgments, abbreviations, spelling errors, and so on. These language phenomena hamper standard text classifiers, which stumble against big number of Out-Of-Vocabulary words. Therefore, the first goal of the text normalization is to improve the feature extraction procedure. For instance, we would prefer to treat all variations of the word “porn” including its abbreviations and forms with spelling errors as a single feature. Secondly, the text normalization improves readability of the output by the term extraction module.

The text normalization is performed by an algorithm which learns rewriting rules from parallel aligned corpora. This normalization is loosely inspired by previous works on SMS normalization (Beaufort et al., 2010). This method shares similarities with both spell checking and machine translation approaches. In our system, all lexicons, language models and sets of rules are compiled into finite-state machines and combined with the input filename by composition, a special operation defined on (weighted) transducers and on (weighted) automata. We use our own finite-state tools: a finite-state machine library and its associated compiler (Beaufort, 2008). In conformance with the format of the library, the compiler builds finite-state machines from weighted rewrite rules, weighted regular expressions and n-gram models. We first tested this approach on the normalization of text messages. The algorithm and its models are described in (Beaufort et al., 2010). In order to learn a normalization model we needed a sequence alignment at the character-level. The best sequence alignment was obtained by applying the algorithm described in (Cougnon and Beaufort, 2009). This algorithm gradually learns the best way of aligning strings.

The filename normalization component is evaluated similarly to the filename classification component. The evaluation is performed on a corpus of aligned filenames by 10-fold cross-validation. The system is trained 10 times, each time leaving out one of the subsets from the training corpus, but using only the omitted subset as test corpus. The normalization system is evaluated in terms of BLEU score, Word Error Rate (WER), and Sentence Error Rate (SER).

3. The First Results

The first experiments with the language analysis module described above were conducted on a dataset of regular pornographic files for three reasons. First, the CSA data is a special case of pornographic content. Therefore, the “standard porn” is useful to develop the system, as well as for initial selection of parameter configurations. In these initial experiments the names of regular pornographic files serve as a substitute for child pornography filenames. Pornography can be expected to share important characteristics with CSA material (like general sex-related vocabulary, types of file extensions, text indicating technical parameters of the media file, etc.). Second, at the time of writing this article, the CSA data were not yet provided by our law enforcement partners due to various administrative and legal issues.

In the experiment described in this article, the system was trained to separate a title of a porn gallery from a title of an encyclopedia. Our training dataset consisted of regular pornographic data crawled from the four specialized porn sites: PicHunter[3], PornoHub[4], RedTube[5], and Xvideos[6]. In particular, each positive training example was composed of a title of a porn video/gallery and tags associated with it. We collected 51350 pornographic titles. We used as the negative training examples 55000 randomly selected titles of the English Wikipedia, each composed of at least 4 words. The full dataset was composed of 106.350 titles. We kept 10% of the data (10635 texts) for the validation. We selected from rest 90% of the data 93.629 titles which are represented with at least two features after all the preprocessing steps. These 93.629 titles were used to train a binary classifier. We did not perform any feature selection – all 39.127 lemmas extracted from the training fold were used as features.

It is important to mention that since the text normalization component was not yet fully integrated in the language analysis module, we used a simplified text normalization procedure in the experiment described here. First, the titles were cleaned up from the numbers and the special symbols. Second, they were POS tagged and lemmatized with TreeTagger (Schmid, 1994). All the standard stopwords (except the “sex-related” ones such as *him*, *her*, *woman*, *man*, etc.) were removed from the titles. Examples of two negative and two positive training examples are provided below:

```
<text class='negative'>
<original>Contractors and General Workers
Trade Union</original>
<lemmas>contractors#NNS#contractor
and#CC#and general#JJ#general
workers#NNS#worker trade#NN#trade
```

[3] http://www.pichunter.com/
[4] http://www.porno-hub.com/
[5] http://www.redtube.com/
[6] http://www.xvideos.com/

```
union#NN#union</lemmas>
</text>

<text class='negative'>
<original>1957-58 American Hockey League
season</original>
<lemmas>1957-58#CD#1957-58
american#JJ#American hockey#NN#hockey
league#NN#league season#NN#season</lemmas>
</text>

<text class='positive'>
<original>Husband catches his wife fucking
with his brother .</original>
<lemmas>husband#NN#husband
catches#VVZ#catch his#PP$#his wife#NN#wife
fucking#VVG#fuck with#IN#with his#PP$#his
brother#NN#brother .#SENT#.</lemmas>
</text>

<text class='positive'>
<original>Slim Can Bearly Take The
Dick .</original>
<lemmas>slim#JJ#slim can#MD#can
bearly#RB#bearly take#VV#take the#DT#the
dick#NN#dick .#SENT#.</lemmas>
</text>
```

Results of our preliminary experiments are presented in Table 1. Our results show that a Support Vector Machine (SVM) or a Regularized Logistic Regression (LR) can distinguish a Wikipedia title from a pornographic video title with accuracy of 96-97%. In particular, the best results were obtained with C-SVM with liner kernel (96.97%). We tested also other kernels of the C-SVM and nu-SVM, but the linear kernel appeared to provide the best results. The training of a model with the linear kernel is also much faster. These results suggest that the titles of encyclopedia are linearly separable from the pornographic titles in the vector space of 39.127 lemmas (therefore the complex kernels are not required). Figure 2 depicts results of the metaparameter optimization of the C-SVM with linear kernel with the grid search. As we can see, this procedure improved the accuracy by 0.37%.

Our results confirm the correctness of the chosen methodology for the filename classification. However, separating CSA from regular porn is much more challenging due to overlapping vocabularies, and because of the lower number of available CSA filenames. Therefore, we expect that the accuracy of the final classifier will be lower. The next stage of our project is training the system on the real CSA data provided by our police collaborators.

4. Conclusion

In this paper we described the principal functions of the language processing module of the iCOP forensics toolkit. Next we presented the architecture of the module and outlined its main components. We listed the language resources used in the filename classifier, the term extractor, and the filename normalizer. Next, this work also presented the first filename classification results with the developed language processing module. Our experiments have shown that the state-of-the-art linear models such as Support Vector Machine or Regularized Logistic Regression are able to distinguish titles of the porn galleries from the titles of encyclopedia articles with an accuracy of 97%. Finally, we discussed the directions for the future research and developments of the module.

Classifier	Training	Accuracy
C-SVM, linear kernel	8m 59s	**96.97%**
C-SVM, polynomial kernel	15m 11s	51.71%
C-SVM, RBF kernel	22m 20s	51.71%
C-SVM, sigmoid kernel	14m 58s	51.71%
nu-SVM, linear kernel	12m 48s	88.20%
nu-SVM, poly. kernel	4m 39s	79.77%
nu-SVM, RBF kernel	26m 49s	88.35%
nu-SVM, sigmoid kernel	14m 5s	87.45%
L2-reg.L2-loss SVM (dual)	0.364s	96.45%
L2-reg.L2-loss SVM (primal)	0.459s	96.52%
L2-reg.L1-loss SVM (dual)	0.366s	96.47%
L1-reg. L2-loss SVM	0.162s	96.47%
L2-reg.L2-loss LR (primal)	0.548s	96.24%
L1-reg. LR	0.388s	93.95%
L2-reg. LR	1.176s	96.27%

Table 1: Results of the title classification with different learning algorithms (default metaparameters were used).

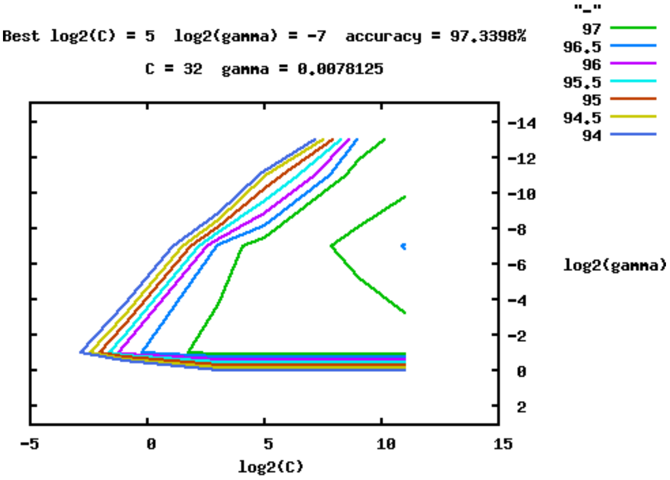


Figure 2: Optimization of the metaparameters of the title classifier (C-SVM, linear kernel) with the grid search.

5. Acknowledgements

This work has been funded by the EC Safer Internet project iCOP. Stéphanie Weiser from Université catholique de Louvain provided comments and remarks, which considerably improved quality of the paper.

6. References

Beaufort R. (2008). Application des Machines à Etats Finis en Synthèse de la Parole. Sélection d'unités non uniformes et Correction orthographique. *Ph.D. Thesis, Faculty of Computer Science, FUNDP, Belgium*.

Beaufort, R. Roekhaut, S. Cougnon, L.-A. Fairon C. (2010). A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *Proceedings of ACL 2010*, pp. 770-779.

Chang C.-C. and Lin C.-J. (2011) LIBSVM: a library for support vector machines. In *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27.

Cougnon, L.-A. Beaufort, R. (2010). SSLD: a French SMS to Standard Language Dictionary. In *Proceedings of eLEX 2009*, pp. 33-42.

Fan, R.-E. Chang, K.-W. Hsieh, C.-J. Wang, X.-R. and Lin C.-J. LIBLINEAR: A library for large linear classification. In *Journal of Machine Learning Research 9*, 1871-1874.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of International Conference on New Methods in Language Processing*. 44–49.

Ugles A., and Stahl, A. (2011). Automatic Detection of Child Pornography using Color Visual Words. In *Proceedings of International Conference on Multimedia and Expo.*

Design of a Controlled Language for Critical Infrastructures Protection

Simona Cantarella,Carlo ferigato,Evans Boateng Owusu
Joint Research Centre of the European Commission
Via E. Fermi, 1 I-21027 Ispra, Italy
E-mail: <simona.cantarella || carlo.ferigato || evans-boateng.owusu>@jrc.ec.europa.eu

Abstract

In this paper, we describe a project for the construction of controlled language for *Critical Infrastructures Protection* (CIP). This project originates from a need to coordinate and categorize the communications on CIP at the European level. These communications can be physically represented by official documents, reports on incidents, informal communications and plain e-mail. We explore the application of traditional *library science* tools for the construction of controlled languages in order to achieve our goal. Our starting point is an analogous work done during the sixties in the field *of nuclear science* known as the *Euratom Thesaurus.*

Keywords: controlled language, thesaurus management, thesaurus display, critical infrastructures protection.

1. Introduction

Controlled languages are as old as the problem of ordering objects in classes. Along the centuries, library science has developed several tools and methods for building, structuring and displaying controlled languages aiming at establishing an artificial communication tool between the indexers and the community that accesses the indexed documents. This artificial language is not fixed once for all; it has its own history and changes according to the changes in the community using it.

Controlled languages for indexing and retrieval can be plainly composed by unstructured *index terms* representing in this way a *taxonomy* for the simple *qualitative measurement* of a document against a measurement grid. By adding structure to this artificial language, index terms can be mutually related establishing in this way a thesaurus. A *thesaurus* with a broad coverage of subjects and endowed by a schematic representation of its index terms is usually called a *classification schema*. Thesaurus construction and maintenance is consequently a communication discipline with its own methods and good practices (Aitchinson et al., 2000). With the construction of a multilingual thesaurus for critical infrastructures protection, we aim at addressing a problem whose importance is growing at the European level: ordering of the communications and of the documents exchanged in the CIP context. Documentation centres on specific critical infrastructures exist since many years. The paradigmatic example is the *Community Documentation Centre on Industrial Risk* established as a consequence of the Seveso Directive (European-Union, 1982). More recently, other documentation centres are growing in different sectors related to CIP as a consequence of the implementation of the *European Programme for Critical Infrastructures Protection* (European-Union, 2007). Moreover, informal communication systems, coordinating the work of the EU Member States in the field of CIP are established.

With the aim of classifying the documents stored and exchanged through these communication systems, we propose a CIP thesaurus as a useful ordering tool. In order to establish a sound method for the construction and maintenance of such a thesaurus, we analysed the work done in the sixties for the construction of a thesaurus in a similar field: the *Euratom thesaurus*. In the following sections we will describe the method and the historical context in which the Euratom Thesaurus was developed by enlightening that the *semantics of index terms are exclusively the indexed documents.*

Subsequently we will describe the display techniques we consider useful for the use of a thesaurus as a controlled language. In the final section we will describe the steps planned for the construction of the CIP thesaurus.

2. The Euratom Thesaurus as a controlled language for a community of workers in the nuclear field

At the end of World War II, nuclear energy was seen both as one of the main future sources of energy in Europe and a weapon whose production should have been kept under control. To this purpose, in 1957, six European countries signed in Rome the Euratom Treaty, formal name for the European Atomic Energy Community.

The Article 8 of Euratom Treaty formally established the JRC, the Joint Nuclear Research Centre. As explicitly stated in the Article 8 , among the duties of the JRC, the development of a European controlled language on the nuclear field was foreseen. This controlled language was called *Euratom Thesaurus*. This thesaurus was developed by CETIS, the *Centre Europèen de Traitmente de l'Information Scientifique*, established at the JRC in Ispra, Italy, in close cooperation with the Document Analysis Group of CID ---*Centre d'Information et Documentation*--- in Brussels.

2.1 The first edition

The Euratom thesaurus, published in two different editions, resulted in an excellent tool for indexing and information retrieval and a qualified controlled language on the subject of nuclear energy. The first edition, published in 1964 (EUR500e, 1964), was based on the analysis of forty thousand documents and it was divided into three related parties:

Alphabetical list of general-purpose keywords.

General glossary of non-keywords terms.

Forty-two terminology charts.

The alphabetical list of general-purpose keywords was formed into:

1,230 general keywords.

1,836 keywords representing inorganic compounds.

1,404 keywords representing nuclides.

The general glossary of *non-keywords* terms was divided into: Synonyms and near-synonyms of thesaurus keywords; proper names used to designate theories and methods, names of projects, reactors, alloys, minerals, zoological species, and so on. It contained more than 2.000 non-keywords terms, together with references to one or more Thesaurus terms that should have been used instead. The references were:

USE, for compulsory assignment of one keywords.

USE. . . +, for compulsory assignment of several key-words.

SEE, for optional assignment of the keywords referred to, if pertinent.

SEE. . . +, for optional assignment of one of the keywords proposed, as far as pertinent.

Finally, the third part of the first edition of Euratom thesaurus was split up into forty-two keyword terminology charts, which corresponded to the subject field within Euratom's fields of interest. Each of these groups represented graphically only the keywords and, within them, the hierarchical and associative relationships between keywords were represented by arrows: the use of the arrows, in this case, replaced the traditional standard references present in other types of thesauri. Normally, the direction of the arrows was from the higher to the lower generic level, while the terms with the same semantic level were connected by two-way arrows. Other arrows linked semantic keywords belonging to different domains of meaning.

2.2 The second edition

The second edition of the Euratom Thesaurus, published in1966, was divided into two related parts:

Dictionary of index terms (EUR500e-I, 1966).

Fifty-seven terminology charts (EUR500e-II, 1967).

The dictionary of index terms, a single alphabetical list of 19,183 words obtained by merging the list of keywords and the glossary of non-keywords terms, was so composed:

Keywords list:

1,199: general-purpose keywords.

1,760: keywords representing inorganic compounds.

1,635: keywords representing nuclides.

71: keywords representing alloys.

Glossary of non-keyword terms:

11,030: accepted terms.

3,488: forbidden terms.

The accepted terms were referred to concepts differing slightly from existing keywords, concepts represented by a combination of keyword terms of inadequate generic levels, proper names used to designate theories and methods, names of projects, reactors, alloys, compounds, minerals and biological species, etc. The forbidden terms, instead, included synonyms or abbreviations of keywords or of *accepted* terms.

In the dictionary of index terms, the forbidden terms were always preceded by a dash, the non-keywords were always followed by a references, such as USE, USE ... +, SEE and SEE... + and the keywords were preceded by the usual notation, which referred to the semantic domains of belonging. The second part of the second edition of Euratom thesaurus was split up into fifty-seven terminology charts: in this edition, each chart represented not only keywords but also the main non-keyword terms and the hierarchical and associative relationships between keywords were represented by links, not arrows as the in first edition.

2.3 Thesaurus updating and maintenance

The primary method of maintenance for the Euratom Thesaurus was to periodically determine the frequency of keyword assignments: low-frequency keywords were eliminated and replaced with more general terms. The position of high-frequency keywords could have been divided between a number of newly introduced terms, representing more specific concepts. Finally, concerning indexing and retrieval, both the indexer and the user used the Thesaurus' alphabetic list as a checklist, noting down every pertinent keyword. Concepts that were not found in the keyword list were looked up in the glossary. If the glossary failed to provide an answer, they used the graphic display schemes. Starting with a relevant term, they simply followed the arrows (or links) leading to other keywords, taking note of all words which were pertinent.

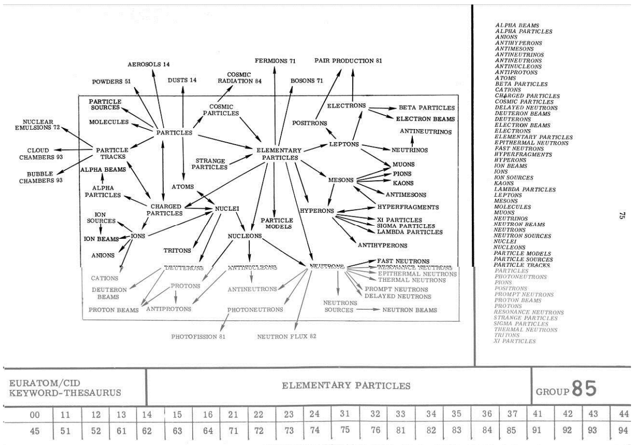


Figure 1: Terminological chart n. 85 from the 1st edition of the Euratom Thesaurus

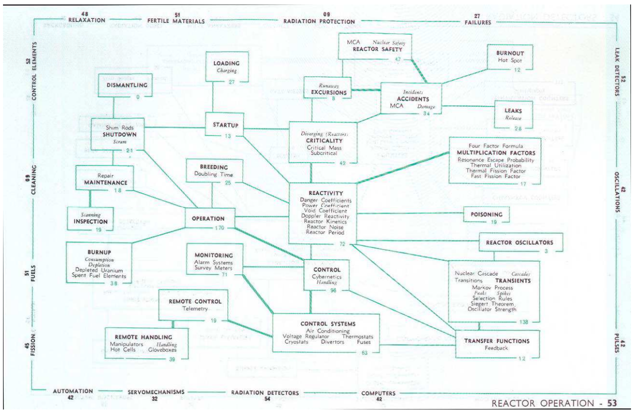

48 RELAXATION
51 FERTILE MATERIALS
09 RADIATION PROTECTION
27 FAILURES
52 CONTROL ELEMENTS
52 LEAK DETECTORS
MCA Nuclear Safety REACTOR SAFETY 47
BURNOUT Hot Spot 12
LOADING Charging 27
DISMANTLING 0
Runaway EXCURSIONS 8
Incidents ACCIDENTS MCA Damage 34
LEAKS Release 28
Shim Rods SHUTDOWN Scram 21
STARTUP 13
Diverging (Reactors) CRITICALITY Critical Mass Subcritical 42
Four Factor Formula MULTIPLICATION FACTORS Resonance Escape Probability Thermal Utilization Thermal Fission Factor Fast Fission Factor 17
09 CLEANING
42 OSCILLATIONS
Repair MAINTENANCE 18
BREEDING Doubling Time 25
REACTIVITY Danger Coefficients Power Coefficient Void Coefficient Doppler Reactivity Reactor Kinetics Reactor Noise Reactor Period 72
Scanning INSPECTION 19
OPERATION 170
POISONING 19
REACTOR OSCILLATORS 3
BURNUP Consumption Depletion Depleted Uranium Spent Fuel Elements 38
MONITORING Alarm Systems Survey Meters 71
CONTROL Cybernetics Handling 96
Nuclear Cascade Cascades Transitions TRANSIENTS Markov Process Peaks Spikes Selection Rules Siegert Theorem Oscillator Strength 138
51 FUELS
REMOTE CONTROL Telemetry 19
CONTROL SYSTEMS Air Conditioning Voltage Regulator Thermostats Cryostats Divertors Fuses 63
TRANSFER FUNCTIONS Feedback 12
45 FISSION
42 PULSES
REMOTE HANDLING Manipulators Handling Hot Cells Gloveboxes 39
AUTOMATION 42
SERVOMECHANISMS 32
RADIATION DETECTORS 54
COMPUTERS 42
REACTOR OPERATION - 53

Figure 2: Terminological chart n. 53 from the 2nd editionof the Euratom Thesaurus.

2.4 The INIS thesaurus

The Euratom Thesaurus is presently incorporated into another specific thesaurus called the ETDE/INIS — International Nuclear Information System /Energy Technology Data Exchange—Joint Thesaurus(INIS, 2012) but the terminological charts have not been included into this thesaurus.

3. Display techniques for controlled languages

A thesaurus exists primarily to coordinate the actions of indexing and searching in the community of its users. It aids in the search strategy by establishing how precisely the indexer can describe the interests of the requester and it aids in indexing by establishing how precisely the indexer can describe the subject matter of the documents. A well-structured, clear and varied thesaurus displays are necessary in helping the indexer to be specific, consistent and exhaustive in their choice of terms and also the user or searcher to easily retrieve documents of interest.

The major standard display types (Lancaster, 1972) are described in the next sections. The following abbreviations are used in the display to describe term details: BT for Broader term, NT for Narrower term, RT for Related term, SN for Scope note, UF for Used for and USE for Use this term instead. The following examples are extracted from the *Anne & Lynn Wheeler Security Taxonomy and Glossary*(Wheeler and Wheeler, 2012).

3.1 Alphabetical display

Terms are arranged in alphabetical order with or without term details. The conventional alphabetical display referred to as Thesaurus Engineering and Scientific terms(TEST) display lists of preferred terms with one level NT, BT hierarchy and other term details.
Example:

Security of data
UF data security
    security of information
NT authorisation
    computer crime
BT security
RT access control
    copy protection
    security of information
USE security of data

The conventional alphabetical display could also have a multilevel hierarchy display. Example:

security target
BT1 target
BT2 security
NT1 functional component
NT2 object
RT security requirements

3.2 Hierarchical display

Only hierarchical relationships among terms are illustrated in this type of display. This type of display is important because it saves the user time. Since all levels of broader and narrower terms are visible immediately there is no need to navigate from term to term to find the full hierarchy. It gives the overview of an entire subject area showing the depth of coverage given to a subject.
Examples:

Security
. assurance
. . effectiveness
. . evidence
. baseline
. . baseline architecture
. component extensibility
. . security target

. . . functional component.
. . . security requirements

Another variant also referred to as the two-way hierarchy lists the broader terms on top of the preferred term and the narrower terms below. Example:

. . security
. assurance
security target
. security requirements
. . security requirements baseline

3.3 Rotated display

Each term is listed multiple times with respect to each non-stop word in the term in alphabetical order. This type of display helps the user to find a multi-word term by looking for any one of its component words.

The Keyword in context(KWIC) display example:

baseline **architecture**
vulnerability **assessment**
assurance
baseline
baseline architecture
component extensibility
functional **component**
computer security
criteria of **control**

The Keyword out of context (KWOC) display lists of all the terms containing a particular keyword. Example:

criteria	criteria of control
	evaluation criteria
security	computer security
	IT security
	security requirements
	security target
	security testing

3.4 Graphical display

This type of display makes it possible to easily draw visual impressions of concepts and relations. Graphical display can communicate relationships among terms more effectively to some users. Examples of graphical display are in figures 1 and 2 above.

Each of the display types discussed above does not stand alone. Used together, they represent the *grammar* of the overall controlled vocabulary scheme.

4. Construction of the CIP Thesaurus

The CIP thesaurus will be constructed along the lines described above: First of all, the corpus of documents giving meaning to the index-terms has to be chosen. To this end, we are considering three distinct corpora of documents:

The principal and secondary European legislation already indexed with a selected sub-set of the Eurovoc (EUROVOC, 2012) thesaurus.

A specific set of journal articles already indexed with a selected subset of the INSPEC thesaurus (IET, 2007).

The documents already recorded in electronic format in the documentation centres described in the introduction above.

Subsequently, a cataloguing process should be performed over the selected corpora. At this point, the selection of index terms and their organization in the standard and graphical display can be performed in agreement with the method used for the construction of the second edition of the Euratom thesaurus. While the target will be the manual multilingual construction of the thesaurus in English, French and German, the process of selection of index terms will be eventually helped by basic tools for the elimination of stop words and frequency counting available for the English language. In any case, the starting point for the collection of index terms will be the alignment of the selected index terms already associated to documents extracted from the thesauri above.

The process of construction of the thesaurus will be performed with the help of tools for multilingual thesaurus construction like *PoolParty* (PoolParty, 2012) or *VocBench* (VocBench, 2012).

In conclusion, we give an example of a bibliographic record of a specific article, in electronic format, taken from IEEE Xplore digital library.

This bibliographic record was realized using the ISBD (ER), the International Standard Bibliographic Description for electronic resources. It specifies the requirements for the description and identification of such items, assigns an order to the elements of the description and specifies a system of punctuation for the description. The primary purpose of the ISBD is to provide the stipulations for compatible descriptive cataloguing worldwide in order to aid the international exchange of bibliographic records between national bibliographic agencies and throughout the international library and information community.

Our job was to add, to this bibliographic record, in addition to the expected fields of ISBD (ER), another area containing the index terms related to the semantics of our article. The index terms were chosen using INSPEC and Eurovoc, the main thesauri of reference. The nature of these indexing tools are different: the first is a specific tool made for a specific domain of interest, the engineering field; the second is a thesaurus built up specifically for processing the documentary information of the EU institutions, so it has a generic framework. Following, the record:

Critical infrastructure protection [electronic resources] : A 21st century challenge / Madjid Merabti, Michael Kennedy, William Hurst. - Electronic data (6 files: 866.731 Bytes). - [S.l.: s.n., 2011]. - In: International Conference on Communications and Information Technology (ICCIT), 2011. - Means of access: World Wide Web. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5762681&tag=1. - Title from first display of information. - ISBN: 9781457704017.

Index terms:

INSPEC: *Controlled Indexing*: computer networks, telecommunication security.

Non controlled Indexing: advanced gaming, critical infrastructure protection, critical national infrastructure, cyber warfare, networked computer systems, targeted attacks, visualization techniques.

EUROVOC:

Preferred terms: digital technology, computer crime, data protection, computer system, systems interconnection, information technology.

Non preferred terms: informatics, data-processing system, computer compatibility, communications technology, computer-related crime, cybercrime, cyber-vandalism, security of information, open systems interconnection.

To assign the index terms, we used both of the thesauri, looking first in their alphabetical display for the name of the specific subject of our interest and then using the structure of the entry to suggest alternative terms. We considered it opportune to indicate both the Controlled Indexing or Preferred terms and the *Non controlled Indexing* or *Non preferred terms*, so as to have many terms to associate.

As can be seen from the above example, in the two thesauri, the term Critical Infrastructure Protection, which is the specific object of the analysed publication, is not among the index terms but it appears only as a non-controlled term. Consequently, it is necessary to develop a controlled language for this sector.

5. References

J. Aitchinson, A. Gilchrist, and D.J.L. Bawden. 2000. *Thesaurus Construction and Use: a Practical Manual*. Aslib, London.

EUR500e-I. 1966. Euratom-thesaurus: indexing term used within Euratoms nuclear documentation system. Luxemburg. European Atomic Energy Community, directorate *Dissemination of Information*, Centre for Information and Documentation-CID.

EUR500e-II 1967.Euratom-thesaurus, part ii: terminology charts used in Euratoms nuclear documentation system. Luxemburg. European Atomic Energy Community, directorate Dissemination of Information, Centre for Information and Documentation-CID.

EUR500e. 1964. Euratom-thesaurus: keywords used within Euratom’s nuclear documentation system. Luxemburg. European Atomic Energy Community, directorate Dissemination of Information, Centre for Information and Documentation-CID.

European-Union. 1982. *Council Directive 82/501/EEC of 24 June 1982 on the major-accident hazards of certain industrial activities.*

European-Union. 2007. *Council Decision of 12 February 2007 establishing for the period 2007 to 2013, as part of General Programme on Security and Safeguarding Liberties, the Specific Programme Prevention, Preparedness and Consequence Management of Terrorism and other Security related risks.*

EUROVOC. 2012. The Eurovoc thesaurus. IET. 2007. *The INSPEC thesaurus*. Institution of Engineering and Technology.

INIS. 2012. The etde/inis joint thesaurus.

F.W. Lancaster. 1972. *Vocabulary Control for Information Retrieval.* Information Resources Press, Washington, D.C.

PoolParty. 2012. The poolparty thesaurus manager.

VocBench. 2012. The vocbench thesaurus manager.

A. Wheeler and L. Wheeler. 2012. Security taxonomy and glossary.

Authorship Identification to Improve Public Security

Aleš Horák, Karel Pala, Jan Rygl
NLP Centre
Faculty of Informatics, Masaryk University
Brno, Czech Republic
E-mail: {hales,pala,xrygl}@fi.muni.cz

Abstract

In the paper, we present details of a new project aimed at automatic web document analysis for the purpose of authorship attribution based on various stylistic and grammatical features of the text. We describe the corresponding system modules with their expected functionality and provide examples of text processing and evaluating techniques.

Keywords: authorship attribution, web document analysis

1. Introduction

The goal of the project is to develop and implement language processing techniques of enabling to analyze linguistic phenomena (texts) whose content is related to the domestic (as well as international) extremism (neonacism, anarchism, racism, terrorism). The research is directed especially to the issues of determining the probability of the authorship of texts. The starting point is a computer analysis of the appropriate data files obtained from the Web and other resources. The attention is paid to all web content produced by the domestic (and later also international) extremistic groups and parties, i.e. web addresses, forums, chats, blogs, social networks and other resources as well.

Approaches and methods developed in the project can contribute to the further advance in the area of security and security informatics and they will facilitate the domestic and international security. We are convinced that the danger resulting from the extremism and terrorism is significant and that it is necessary to fight both extremities in the cyberspace, if we do not want to concede that a lot of young people would be infected, radicalized and recruited by means of the Web to endanger democracy and peace.

The main result of the project will be formed by a compact modular system whose development will include several parts connected in a natural way. The individual developed modules will be improved and extended in parallel. In the first phase of the project it is necessary to collect from the Web and other suitable resources as much text data from the mentioned areas as possible and store them in the form of text corpora related to extremism.

To this end, the tools for obtaining text from the Web are being created, namely specialized web document crawlers. The obtained data are semiautomatically annotated with expert guidance. In this way we will get training and testing data for the development of the necessary automatic techniques. From our existing experience it follows that the success rate of most of the automatic methods heavily depends on the size of the training data that are at hand. Since at the beginning the size of the testing data are not large enough, we have created a large corpus of the standard reference texts that is used for comparison with the investigated documents. While in English we can rely on frequented phenomena, in Czech we have to stand up the data sparseness problem, thus we are looking for new techniques based on the contrastive analysis of the reference texts.

2. System Architecture

Following the methodology described in the previous section, we are dealing with the development of the modules, which are directly related to the text authorship.

2.1 Fact Extraction

One of the modules performs an extraction of the basic facts, especially temporal and space ones, from utterances as they are related to authors. We assume that these data can be used by experts as a background for determining authorship, as it is demonstrated by the example in Figure 1.

2.2 Intelligent Web Monitoring

Further modules determine probability that two documents were produced by the same author. At the same time as documents can be used e.g. longer articles, collections of the contributions to a discussion written under one login or just short text fragments.

An intelligent exploitation of the documents retrieved from Internet needs a description of the format of the stored document meta-data. Since it is tedious to manually extract the structure of web pages to be able to download information about documents and authors, we employ a new, semi-automatic approach. It consists of the following 4 steps:

Internet forum:
Sall

> I will add a contribution about antifa. In the course of the blocade of nazi march at the anniversary of Crystal Night I was happy that they (antifa) were on my side. Then (around 18:00) when the nazi group understood that it does not make sense and pulled out of the Law Faculty I don't know where. But I wanted to say that it split there in peace, so I and my friends were glad that we expressed our view peacefully and went home. At home I then could watch on TV how antifa fought police and thus they attracted attention as the ones who picked a fight and were making ruckus. To tell the truth, I was a bit sad about this.

Extracted facts:

Who:	**What:**	**Where:**	**When:**	**With whom:**
Sall	was	In the course of the blocade of nazi march at the anniversary of Crystal Night		
they (antifa)	were on my side			Sall
the nazi group	pulled out		around 18:00	

Figure 1: Example of fact extraction

1. Firstly, a domain from which documents are going to be extracted is selected and visited by an operator. The operator manually registers to the domain using data describing her institution. Then it is necessary to submit a small number of documents $d_1, \ldots, d_k$ (e.g. discussion posts, blogs) while logged as the registered user.

2. In the second step, a web downloader is used to obtain web pages $P_1, P_2, \ldots$ in the domain until all pages containing information about documents $d_1, \ldots, d_k$ are found: $P_{d1}, \ldots, P_{dk}$.

3. In the third step, the HTML tree structure of the selected pages is detected by a HTML parser. Then minimal sequences of HTML tags are extracted to describe each attribute of the documents $d_1, \ldots, d_k$ using local search heuristics. It is important that each sequence describes the same information in all downloaded web pages, e.g. the title sequence defines a path to information about the title for every document. The structure of the domain is stored into the database as generated sequences of HTML tags.

4. Finally, all documents from the domain are downloaded by the crawler and processed. With the knowledge of the web page's structure, only data relevant for the authorship identification are collected and saved in the database. The algorithm is summarized in Table 1.

In any future attempt to retrieve documents from already processed domain, $P_{d1}, \ldots, P_{dk}$ are downloaded again and their content is compared to the saved data in the database ($d_1, \ldots, d_k$). If the content differs, either documents were edited (which is unlikely because documents were created by the operator), or the structure of pages in the domain was changed. Therefore, in this case all 4 steps are executed again. Otherwise, only new pages are processed. Further evaluation of this approach can be found in (Rygl and Horák, 2011).

2.3 Authorship Attribution

Other algorithms will determine authorship probability on the ground of the more extensive database of the annotated documents.

Clustering of anonymous documents according to the authorship is very difficult. There even do not exist any recommended metrics for measuring the quality of a particular clustering in the authorship identification problem. We conducted some experiments but the results' accuracy was low. Although similarity of two documents can be compared with relatively high accuracy, for creation of large clusters many comparisons are made and even marginal errors decrease total accuracy significantly.

Therefore anonymous documents are not clustered and only documents signed by authors are put together. In

action	example
Select domain	*D = www.domain.com*
and register as author	*(Name, E-mail) → D*
and submit article	*(Title, Text, Name) → D*
Download domain texts	*documents* $t_1, \ldots, t_n \in D$
Search inserted document	t_k *= (Title, E-mail, Name)*
Extract structure of document	*Title*$_k$*:body/div[@content]/h3*
	Text$_k$*:body/div[@content]/p*
	Author$_k$*:head/title*
Process downloaded documents	*Title*$_k$*:Introduction post*
	Text$_k$*:Text about web page's topic…*
	Author$_k$*:NLP Center*

Table 1: Domain structure identification

order to adapt to data from an online environment for which identification of authors are not unique, an operation merging two clusters is allowed. It is very important to process data from different domains because the author's accounts may vary. Either it is a cosmetic change of identity (e.g. size of letters, leave out one word), or the author uses a completely different pseudonym.

Whenever a new author is inserted into the database, he or she is compared to each known author. If two authors' documents differ only marginally, their identities are connected. On the contrary, authors with same identities from different web domains are not linked automatically, their similarity has to exceed a specified limit that is more tolerant than in the case of two different identities. This is necessary because many pseudonyms and names overlap. Despite the fact that the operation is time consuming, it is affordable because it is sufficient to apply it only to authors of documents with a similar theme and each author is processed only once.

The authorship detection is improved if we replace authors' similarities to the document (according to characteristics) by author's positions in rankings (generated from these similarities).

For example, we are given a set of three authors A_1, A_2, A_3, unknown document d and the characteristic C with values $C(d, A_1) = 0.5$, $C(d, A_2) = 0.7$, and $C(d, A_3) = 0.2$. The corresponding ranking function R can be given as $R(d, A_1) = 2$, $R(d, A_2) = 1$, and $R(d, A_3) = 3$. Those values are then used as an input for the machine learning process.

If we consider the problem of authorship identification as a competition among potential authors, we can use a sport analogy: If athletes compete under the same weather, health conditions and the track is always the same, we recognize the best athlete by her score (time, points, etc.). But the best athlete is not necessarily the holder of the best score. What matter is the position of athletes in rankings. It compensates for unequal conditions at each competition. To explain why we need to consider unequal conditions we need two example sets of documents.

Let us have a set of documents d_i written about the same topic by authors A_1, A_2, A_3. One of the documents (document d) of an unknown author is in this set of documents. Since the documents share one topic, they contain many similar words. This fact affects the calculated characteristics (scores are increased), e.g. $C(d, A_1) = 0.8$, $C(d, A_2) = 0.7$, and $C(d, A_3) = 0.9$.

The second set contains documents of different length and topic (we use the same 3 authors as in the set one). Most of the characteristics depend on the statistics of various phenomena in the text. If the documents are of various length and topic, the phenomena occur in varying degrees, reducing the overall scores e.g. to $C(d, A_1) = 0.4$, $C(d, A_2) = 0.3$, and $C(d, A_3) = 0.5$.

The problem is that values 0.9 and 0.5 indicate the same authorship and values 0.7, 0.8, 0.4 and 0.3 are used for the different authorship. The machine learning can deal with the problem at the cost of reduced accuracy.

To optimize the machine learning, we have changed the ranking function R to inverse function $S = 1/R$, therefore, we obtain values 1, 1/2, 1/3, …, $1/N$ instead of values 1, 2, 3, …, N. The main advantage of the featured function S is that there is not such a big difference between problems with different number of authors. The best authors are evaluated equally and the worst authors' values differ absolutely by a small number (compared to the difference between the first positions). This means that learning examples of a problem containing 5 authors are applicable to a problem containing 20 authors.

The suggested function S is consistent across different situations and returns values in the interval $\langle 0, 1 \rangle$, which is recommended for the implementation of Support Vector Machines algorithm we used to process data (Hsu et al., 2010). More detailed evaluation of this approach can be found in (Rygl and Horák, 2012).

3. Conclusion

In the phases of the project we will pay attention to finding authorship in text using special techniques, for instance Writeprint, which serves to the unambiguous identification of the anonymous authors and is based on the signatures associated with with the messages at chats and forums. We are going to extend lexical and syntactic procedures of the traditional authorship analysis in such a way that we will capture the relevant system features (font size, colours, web links) and semantic features of the extremism domain in online texts. Estimated number of the features can be up to hundreds, they will automatically searched for in the Internet messages. Using Writeprint technique for English may lead to relatively interesting results. The Writeprint will be also used for looking up the texts of the given author on Internet. We also will pay attention to the analysis of the expressions denoting emotions. In the first place we will work with the Czech texts, in the following phases of the project we will be interested also in English and eventually other languages (Arabic, Russian)..

4. Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013, by the Ministry of the Interior of CR within the project VF20102014003, by EC FP7 project ICT-248307 and by the Czech Science Foundation under the project P401/10/0792.

5. References

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin (2010). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, Taipei 106, Taiwan. URL:http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Rygl, J. and Horák, A. (2011). A Framework for Authorship Identification in the Internet Environment. In A. Horák and P. Rychlý, editors, *Proceedings of RASLAN 2011*, pages 117-124, Brno, Czech Republic. Tribun EU.

Rygl, J. and Horák, A. (2012). Similarity Ranking as an Attribute for Machine Learning Approach to Authorship Identification. In *Proceedings of LREC 2012*, pages 1-4. accepted for publication.

Unexpected Factual Associations Mining

Wiesław Lubaszewski, Michał Korzycki

AGH University of Science and Technology
Al. Mickiewicza 30, Kraków, Poland
lubaszew@agh.edu.pl, korzycki@agh.edu.pl

Abstract

The paper describes the LSA (Latent Semantic Analysis) algorithm as a tool for mining unexpected factual associations from text corpora. Due to the fact that LSA performs well on text corpora built from short texts it can be a useful tool to analyse e-mails stored in the mail box, chats logs or Internet fora content. Therefore the LSA may serve as a tool in forensic or security analysis.

Keywords: LSA, text mining

1. Introduction

There are cases in which one cannot use structured information patterns, due to the fact that the information to be extracted is not precisely known. In such a situation one can use short texts or text samples as an information pattern to look for texts, which bring similar information in a particular text corpus, e.g. e-mail server resources, server logs, etc.

We use Latent Semantics Analysis (LSA) as the model for text similarity extraction. The tool will accept any user provided text and will return a number of texts selected from the text corpus. All returned texts contain information associated to the information of the input text. The paper discusses and interprets several examples in terms of semantic and factual stereotypes and associations. The described examples and mechanisms are a part of a tool for associative information retrieval used by forensic analysts.

2. The LSA Algorithm

LSA is a word/document matrix rank reduction algorithm, which extracts word co-occurrences in the frame of a text. As a result each word in the corpus is related to all co-occurring words and all texts in which it occurs. This makes a base for an associative text comparison. But due to its statistical nature LSA is unable to distinguish co-occurrences, which are corpus independent semantic dependencies (elements of a semantic prototype) from co-occurrences, which are corpus dependent factual dependencies. This fault brings an unexpected value – the LSA can associate facts described in texts in a way, which may by omitted (rejected) by human reasoning based on semantic and factual stereotypes and driven by *the principle of the least effort*. This ability makes the LSA algorithm a valuable tool in man-made analytics – an analyst can find in the LSA output unexpected or unpredictable factual associations.

3. An associative power of LSA

To demonstrate this LSA ability we need an example. Let’s take the input text:

On Saturday, the Polish Siamese twins Weronika and Wiktoria flew to Poland on a flight from Newark to Krakow – as reported by LOT (Polish Airlines) to the section of PAP (Polish Press Agency) from the suburban New York airport. The two Siamese twins, Weronika and Wiktoria were born on May 26, 1999, in the hospital of the Academy of Medicine in Lublin. They have been in Philadelphia since August 16. They survived a successful operation to separate them in November. The first date to release the two from the hospital was February 11. It was later extended for a week due to Weronika’s fever.

If we use the analyzed text as the input to a SVM algorithm and then, if we perform a search through the PAP corpus, we will find that the SVM returns only texts strongly related to an operation performed on Siamese twins (Figiel A. 2009).

If we use the same text as the input to the LSA algorithm and then, if we perform a search through PAP corpus (45 000 agency news from 2000 to 2003), we will find that the LSA return more than hundred texts, which informational similarity to the input text varies from 1 to 0.3 (Lubaszewski W. and Korzycki M. 2011). This means that the LSA has a much lower precision and much higher recall than the SVM. But the use of classic precision – recall measures would not explain the associative power of the LSA. To analyze the LSA output we must start with a semantic analysis of the content of the input text.

First, one has to distinguish events and objects in the text. We can see the main event (medical treatment) and secondary events, which are: a travel (going back home) and a news broadcast.. Both events are related to the main object (Siamese twins). We can also find secondary objects related to the main event, ie. hospital, academic institution, nationality (Polish) and secondary objects related to the secondary event, ie. airport, airlines and press agency. Now we can look at the LSA output ordered by the similarity to the input text:

a) The first in the output is input text – as expected – the most similar to itself: similarity 1;

b) The next in similarity rank is a text about the further treatment of twins mentioned in the input text – a text ID in parentheses:

Weronika, one of two Siamese twins from Stalowa Wola, was operated on at the University Children's Hospital in Krakow. A plastic plate, which served as a substitute sternum for a year and a half, after an operation to separate her from her sister, was removed from her chest cage. "The operation was a success. Now the problem is, that Weronika is so lively, she'll have to be constantly monitored, because otherwise she'll wiggle free from her bed." said Krystyna Paleń, the child's mother. „She's a very strong child. Nobody suspects any complications after the operation. Everything went well.". The Siamese twins Weronika and Wiktoria, were born in May, 2009, at the Lublin Clinic of Periantology and Obstretics. The girls were joined at the chest and stomach. They were separated at Children's Hospital in Philadelphia in November, 1999.

similarity: 0.552321 (text ID 48363)

c) The third in rank (similarity higher or equal to 0.479) is a group of texts, each of which matches the main event, but may not match the main object. For example:

The French actor Gerard Depardieu, has had a bypass operation in a suburban Paris hospital – as reported by the doctor who performed the operation.

similarity: 0.527882 (text ID 11933)

or the text:

A spokesman for a hospital in Manchester, where the operation to separate the Siamese twin sisters took place, informed that the weaker of the two girls – Mary – did not survive the operation. The second of the two twins – Jodie – is alive. After a complicated operation lasting almost twenty hours, which was concluded Tuesday morning, her condition is listed as "critical but stable".

similarity: 0.521147 (text ID 22759)

A more in depth further analysis of the results show that there is a variety of objects related to main event: baby tiger (25464), a named individual (011933), the Queen Mother (47668), a boy (41698, 487148), an old woman (36890), Siamese twins (22759, 36611)(see Lubaszewski W. and Korzycki M. 2011).

In general, these examples show that the LSA mechanism is able do distinguish the main event and that matching the main event is more important than the matching of the main object.

d) the fourth in rank (similarity 0.479 - 0.471) are texts, which do match the main object, ie. Siamese twins but do not match the main event,

One of two nine month-old brothers who are Siamese twins, who were separated by doctors at the end of June in Krakow, died Wednesday evening as result of blood circulation failure – Dr. Adam Bysiek informed PAP.

similarity: 0.479452 (text ID 12192)

e) the fifth in rank (similarity below 0.47) are texts, which match a secondary event, ie. travel (back home), and where object matching is not so important,

On Saturday, Captain Krzysztof Baranowski returned do Łeba, from where he set out one year ago, aboard the yacht "Lady B". He completed a solo cruise around the world at Villamoura on September 30. He is the first Polish sailor to circumnavigate the world, solo, twice.

similarity: 0.464156 (text ID 19200)

f) the sixth in rank are texts, which match the main event (successful medical treatment), but the object of treatment is unmentioned,

A complicated operation to reattach a hand that was cut off by a buzz-saw at a lumber mill was successfully performed at the Clinic of General and Hand Surgery of the Academy of Medicine in Szczecin.

similarity: 0.458652 (text ID 19278)

g) the seventh in rank are associated texts, that is texts, which do not match the events of the input text (neither main nor secondary), but do match an event semantically associated with the events of the input text - these texts have a similarity measure above 0.3.

In the LSA output we can find texts associated with the main event, for example text, which brings information about an assault:

A swarm of bees has attacked more than 300 people in Pakistan, including many doctors who were participating in a charity march, which was being held to collect money for a hospital.

similarity: 0.338710 (text ID 3608)

The LSA has also associated texts, which bring information about death:

An 88 year-old resident of China, whose death was confirmed by a doctor, awoke after having spent two hours in a casket.

similarity: 0.338157 (text ID 2575)

There are also texts in the LSA output, which are associated with the secondary event of the input text and the secondary object, for example a text associated on the base of semantic opposition: travel (going back home) versus travel (leaving home).

A Polish expedition to study glaciers will set out for Spitzbergen at the end of March – Dr. Marek Górski from the Polar Research Section of the Institute of Geophysics at the PAN (Polish Academy of Sciences) has reported.

similarity: 0.337679 (text ID 2125)

The naive analysis of the LSA output would show that all texts except those with similarity rank (e) and (g) are related to surgery and medical treatment. From this point of view we can say that each text, which do not relate to a surgery brings an unexpected semantic or factual associations. If we look at the result of the naïve analysis from the key-word search perspective we can intuitively suppose that that person, who looks for information related to the input text would not use such key-words as travel, expedition, attack, bees or false diagnose, which are important to find texts with similarity rank (e) and (g) associated by the LSA algorithm. This intuition is supported by an experiment.

4. Man-made Associations

The Participants of the experiment were asked to assign an unlimited number of keywords to the input text and separately to each text, which were classified as an unexpected association. In the experiment separately separately two groups of participants took part. The first group consists of postgraduate students well trained in text analysis. The second group consists of postgraduate students without such training – so called naïve analysts.

Siamese twins	**Analysts assigned keywords**		
keywords	**trained**	**untrained**	**TOTAL**
Siamese	92.31%	100.00%	96.30%
twins	100.00%	85.71%	92.59%
operation	76.92%	50.00%	62.96%
separation	38.46%	64.29%	51.85%
Philadelphia	15.38%	28.57%	22.22%
Weronika	0.00%	35.71%	18.52%
Wiktoria	0.00%	35.71%	18.52%
arrival	7.69%	21.43%	14.81%
successful	23.08%	0.00%	11.11%
Polish	15.38%	0.00%	7.41%

Table 1. Analyst keyword assignments for the main text

If we look at table 1. we can observe that both group of participants almost perfectly assigned keywords, which identify the main event of the text and which also can identify the semantic stereotype behind the main event – there are the most frequent keywords. Only some participants assigned keywords, which identify a secondary event. Finally, we can observe that many (some) participants intuitively tended to restrict the search range by introducing keywords related to a particular real event described in the text, that is proper names, etc. If we take a careful look at the keyword list assigned to the input text we can find only a very small percentage of participants that used the keyword *Polish*, which would match texts about Polish expeditions **(text ID 19200, text ID 2125),** but there is no keyword, which would match the text about a medical error in China **(text ID 2575)** or the text about bees attack in Pakistan **(text ID 3608)**. Keywords assignments less frequent then 7% were omitted.

The same behavioural pattern can be observed in the description of texts, which were associated by LSA.

Swarm of bees	**Analysts assigned keywords**		
keywords	**trained**	**untrained**	**TOTAL**
attack	61.54%	80.00%	71.43%
bee	69.23%	73.33%	71.43%
swarm	69.23%	66.67%	67.86%
Pakistan	38.46%	80.00%	60.71%
marsh	76.92%	40.00%	57.14%
charity	46.15%	26.67%	35.71%

Table 2. Analyst keyword assignments for text id 3608

Medical error	**Analysts assigned keywords**		
keywords	**trained**	**untrained**	**TOTAL**
death	91.67%	60.00%	74.07%
awakening	50.00%	73.33%	62.96%
casket	58.33%	66.67%	62.96%
China	33.33%	73.33%	55.56%
fault	16.67%	13.33%	14.81%

Table 3. Analyst keyword assignments for text id 2575

Spitsbergen expedition	**Analysts assigned keywords**		
keywords	**trained**	**untrained**	**TOTAL**
expedition	100.00%	73.33%	85.71%
Spitsbergen	76.92%	86.67%	82.14%
glaciological	46.15%	86.67%	67.86%
Polish	30.77%	33.33%	32.14%

Table 4. Analyst keyword assignments for text id 2125

To explain this phenomenon one must reach for psychological experiments, which started almost hundred years ago. The word association experiments – we refer to - show that human being most frequently associates words, which form semantic stereotype, e.g. *table – chair* and less frequently words related by factual coincidence, e.g. *table – aunt*. (G. A. Miller and P. N. Johnson-Laird 1976). This language independent observation may be related to all human activity, which is mainly based on semantic stereotypes what is called *the principle of the least effort* (G. K. Zipf 1949). Therefore to associate text, classified as an unexpected association, for example:

A swarm of bees has attacked more than 300 people in Pakistan, including many doctors who were participating in a charity march, which was being held to collect money for a hospital.

similarity: 0.338710 (text ID 3608)

one must go beyond the semantic stereotype, hidden behind the main event of the input text, to start reasoning based on a semantic relation chain, for example the chain: an assault may cause a wound ← a wound may need treatment ← a doctor orders a treatment. But we must stress that a psychological probability that a person concentrated on a specific information is able to perform such reasoning without some external stimulus is very low. We shall argue that the events experienced by a forensic analyst would follow this rule, and therefore the mechanic associations made by LSA may bring value in criminal or security investigation showing a broader set of potential associations than could be looked for by a human analyst.

5. Conclusions

There is no doubt that an associative mechanism of the LSA should be a subject of further studies. But we believe that even the results described in this paper prove beyond doubt that in some cases LSA may serve as a useful tool in forensic or security investigations. That is why an LSA based search tool has been included in a forensic toolkit, as a complementary high recall solution to a high precision script guided search engine.

6. Acknowledgements

The described examples and mechanisms are a part of a tool for associative information retrieval developed for forensic and security analysts in the INDECT project, supported by European Commission under the grant FP7-218086.

7. References

Figiel A., (2009) "The Text as Informational Pattern - Topic Similarity Evaluation by the LSA", In Lubaszewski W. ed, *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu*, Kraków, AGH Press, 2009, original text in Polish

Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988), "Using latent semantic analysis to improve information retrieval." In: *Proceedings of CHI'88: Conference on Human Factors in Computing*, New York: ACM, 281-285.

Landauer, T. K., & Dumais, S. T. (1996). How come you know so much? From practical problem to theory. In D. Hermann, C. McEvoy, M. Johnson, & P. Hertel (Eds.), *Basic and applied memory: Memory in context*. Mahwah, NJ: L. Erlbaum, 105-126.

Lubaszewski W. and Korzycki M, *System for Enhanced Search: A Tool for Associative Information Retrieva*l, Technical Report D9.15., INDECT Project, FP7-218086, Kraków, 2011.

Miller G. A. and Johnson-Laird P.N., Language and Perception, Cambridge, Cambridge University Press, 1976.

Zipf G. K., *Human Behaviour an the Princple of Least Effort*, Cambridge Mass. Addison-Wesley, 1949.Castor, A., Pollux, L.E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1), pp. 37--53.

Precedes: A Semantic Relation in FrameNet

Miriam R. L. Petruck and Gerard de Melo
International Computer Science Institute
Berkeley, California, USA
miriamp@icsi.berkeley.edu, demelo@icsi.berkeley.edu

Abstract

Automatic language processing systems depend on, among others factors, the effectiveness in modeling human cognitive abilities, including the capacity to draw inferences about prototypical or expected sequences of events and their temporal order. Appropriate response to a crisis is as important for public security as are efforts to prevent any such natural or man made disaster. Recent research (Mehrota et al. 2008) has recognized the need for accurate and actionable situation awareness during emergencies, where timely status updates are critical for effective crisis management. The present paper constitutes a contribution to situation awareness for Natural Language Processing (NLP) applications to improve communication among first responders, and features the frame-to-frame semantic relation **Precedes**, as implemented in FrameNet (http://framenet.icsi.berkeley.edu). Specifically, this work demonstrates the necessity and importance of the information encoded with **Precedes** for NLP applications, advocating the inclusion of such information in systems for security applications.

Keywords: FrameNet, semantic relations, inferencing

1. Introduction

The success of any automatic language processing system depends on, among others factors, its effectiveness in modeling human cognitive abilities, including the capacity to draw inferences about prototypical or expected sequences of events and their temporal order. Appropriate response to a crisis is as important for public security as are efforts to prevent any such natural or man made disaster. Recent research (Mehrota et al. 2008) has recognized the need for accurate and actionable situation awareness during emergencies, where timely status updates are critical for effective crisis management. The present paper constitutes a contribution to situation awareness for Natural Language Processing (NLP) applications to improve communication among first responders, and features the frame-to-frame semantic relation **Precedes**, as implemented in the FrameNet database (http://framenet.icsi.berkeley.edu).

Based on the principles of Frame Semantics (e.g. Fillmore, 1975; Fillmore, 1977; Fillmore, 1985) and committed to documenting its findings through corpus attestations,[1] FrameNet is an ongoing lexical resource development project that documents the **valences**, or semantic and syntactic combinatorial possibilities, of English vocabulary. With computer-assisted annotation of examples and automatic tabulation and display of the results, FrameNet records the valences for each word in its lexicon. The project's main product is its lexical database, currently containing over 1,100 frames, nearly 12,500 lexical units, and approximately 190,000 examples. Befitting of the larger endeavor, the database includes frame-to-frame relations for its hierarchy of semantic frames.

Much of the written work and public presentations of the FrameNet team and its affiliates have attended to the most important frame-to-frame relations recorded in the database, **Inheritance**, **Subframe**, and **Using** (Fillmore and Baker, 2004; Fillmore et al., 2004). Petruck et al. (2004) characterized the lexicographic imperative for adding **Inchoative_of** and **Causative_of** to the inventory of semantic relations that FrameNet records. Chang et al. (2002) provided a structured event formalism that translates FrameNet's informal descriptions into a representation appropriate for simulative inference, also offering a means of handling linguistic focus, which FrameNet captures with **Perspective_on**.[2] The present paper focuses on **Precedes**, the semantic relation in FrameNet that captures the notion of temporal ordering, and demonstrates its necessity for natural language understanding (e.g. Buchardt et al., 2009; Shen and Lapata, 2007; Fillmore and Baker, 2001).

2. Frame Semantics

At the heart of Frame Semantics is the **semantic frame**, a schematic representation of an event, object, situation or state of affairs whose **frame elements** (**FE**s) identify participants and props and whose underlying conceptual structure speakers access for both encoding and decoding purposes. The semantic frame, parts of which are indexed by words that *evoke* the frame, is a cognitive structuring device used in the service of understanding (Fillmore, 1985). FrameNet distinguishes three categories of FEs: **core**, **peripheral**, and **extra-thematic**. Core FEs are frame specific and uniquely define a frame, capturing conceptually necessary aspects of the scene. Peripheral FEs identify characteristics of situations and events more generally, including the time or place of

[1] FrameNet primarily uses the British National Corpus, and to some extent the American National Corpus.

[2] Chang et al. (2002) foreshadowed the introduction of **Perspective_on** into FN. **Perspective_on**, a refinement of **Using**, first appeared in data release 1.3 (2006).

an event, as well as the manner in which an event occurs. Extra-thematic FEs situate an event or state of affairs against the backdrop of another event or state of affairs, such as the frequency with which an event occurs, or a description of a participant in an event in terms unrelated to the event. Conceptually, extra-thematic FEs are not part of the frame in which that type of FE appears, instead belonging to more abstract frames where they fill argument roles of their own.[3] A Frame Semantics description of a word identifies the frame or frames that constitute the conceptual basis of a given sense, and specifies the ways that structures headed by the word realize those FEs.

FrameNet defines `Avoiding` as a situation in which an AGENT avoids an UNDESIRABLE_SITUATION under certain CIRCUMSTANCES, where that situation may be an event or an activity.[4] Whereas AGENT and UNDESIRABLE_SITUATION are core Frame Elements, CIRCUMSTANCES is a peripheral FE, since it characterizes an aspect of a wide range of events in addition to avoiding. Following Cruse (1986), FrameNet adopts the **lexical unit** (**LU**) as the focus for lexicographic annotation, defining an LU as a pairing of a lemma and a frame. Among the LUs that figure in `Avoiding` are *avoid*, *avoidance*, *evade*, and *evasion*. Example (1) illustrates the LU *avoid*.v, which evokes the `Avoiding` frame, annotated with respect to that target LU, along with the triples of information that FrameNet records for each FE, including phrase type (PT) and grammatical function (GF).

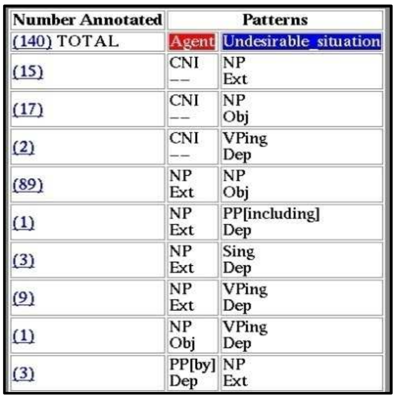
1. [$_{\text{AGENT}}$ The reporter $_{\text{NP/External}}$] **AVOIDED** [$_{\text{UNDESIRABLE\_SITUATION}}$ entering the roped off area $_{\text{VPing/Dep}}$].

The NP *the reporter*, instantiating the AGENT has the GF External; and the VP*ing entering the roped off area* realizes the UNDESIRABLE_SITUATION, functioning grammatically as a Dependent (Dep).[5] A FrameNet lexical entry provides a table of the valence patterns, or combinatorial possibilities, specifying the mapping of semantic roles to syntactic structures and showing the full array of syntagmatic relations for that word. Below, Figure 1 shows a partial valence table for *avoid*.v in `Avoiding`, displaying only the core FEs AGENT and UNDESIRABLE_SITUATION.

Example (2) illustrates FrameNet annotation for a sentence that also realizes non-core FEs, here TIME and CIRCUMSTANCES, the former as PP/Dep and the latter as Sinterrogative/Dep (PT/GF).

2. [$_{\text{TIME}}$ At the beginning $_{\text{PP/Dep}}$] [$_{\text{AGENT}}$ the reporter $_{\text{NP/External}}$] **AVOIDED** [$_{\text{Undesirable\_situation}}$ entering the roped off area $_{\text{VPing/Dep}}$] [$_{\text{Circumstances}}$ while looking for evidence $_{\text{Sinterrogative/Dep}}$].

FrameNet distributes the valence tables for each lexical entry in XML, making this syntagmatic information accessible for use in NLP applications.



Number Annotated	Patterns	
(140) TOTAL	Agent	Undesirable_situation
(15)	CNI --	NP Ext
(17)	CNI --	NP Obj
(2)	CNI --	VPing Dep
(89)	NP Ext	NP Obj
(1)	NP Ext	PP[including] Dep
(3)	NP Ext	Sing Dep
(9)	NP Ext	VPing Dep
(1)	NP Obj	VPing Dep
(3)	PP[by] Dep	NP Ext

Figure 1: Partial Valence Table for `Avoiding`.*avoid*.v

3. Frame-to-Frame Relations in FrameNet

FrameNet records frame-to-frame relations in its hierarchy of semantically organized frames, also making that information available for natural language processing applications. Figure 2 depicts the relevant frame-to-frame relations that FrameNet has recorded for the `Employment_scenario` frame.

Inheritance exists between a parent frame and a child frame under specific circumstances: for each FE, frame relation, and semantic characteristic in the parent, the same or a more specific corresponding entity in the child exists, as in the relationship between `Employment_end` and `Firing`. **Using** is a relationship between a child frame and parent frame in which only some of the FEs in the parent have a corresponding entity in the child; if such exist, they are more specific. Using holds between `Fields` and `Employment_scenario`, where the FEs ACTIVITY, PRACTITIONER and WORK in `Fields` are the more specific instances of TASK, EMPLOYEE and POSITION in `Employment_scenario`, respectively. FrameNet uses **Perspective_on** (Chang et al. 2002) to distinguish between neutral and *perspectivized* frames, the latter identifying different points of view of other participants in the larger scenario. As a consequence, whereas `Employment_scenario` is neutral in terms of participant point of view, `Employer’s_scenario` and `Employee’s_scenario` capture the perspective of the employer and employee, respectively. The **Subframe** relationship characterizes the different parts

[3] See Chapter 3, Ruppenhofer et al. (2010), for a detailed discussion of FEs and coreness.

[4] The paper uses the following typographical conventions: **bold-face** for special terms in the prose; small caps for FEs; bold-face caps for targets; and Courier New for frame names.

[5] FrameNet uses a limited set of grammatical functions. See Chapter 5, Ruppenhofer et al. (2010) for more information.

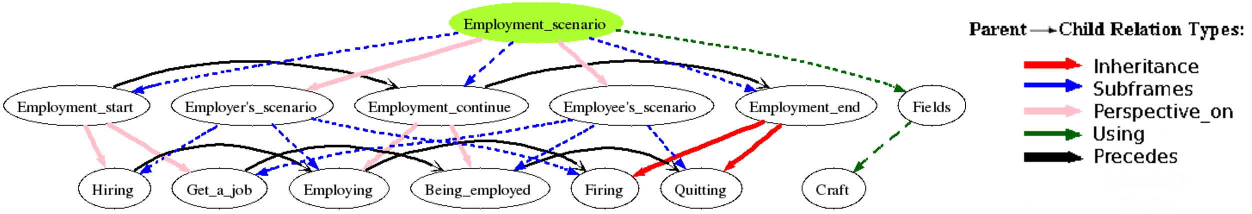


Figure 2: The Employment_scenario and its Frame-to-frame Relations

of a complex event in terms of the sequences of states of affairs and transitions between them, each also describable as a frame. **Precedes** captures the temporal ordering of subevents within a complex event. The relation holds between component subframes of a single complex frame, and provides additional information to the set of **Subframe** relations, as in `Hiring`, `Employing`, and `Firing`, each a separate frame in the complex `Employer's_scenario`. **Precedes** is also the only relation that allows cycles, as for example with repeated hirings, employings, and firings.[6]

4. Fire Fighting: Search and Rescue

Fire fighters are among the first to appear on the scene of an emergency, as occurred on September 11, 2001. Here, we generally characterize the complex Fire Fighting (FF) scenario and then focus on Search and Rescue, a sub-phase of FF, to demonstrate the necessity and importance of **Precedes** for NLP applications.

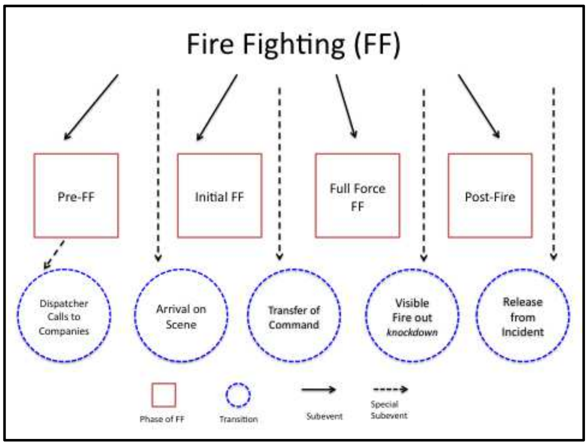


Figure 3: Fire Fighting Phases and Transitions

Figure 3 depicts the major phases and transitions in FF, where FrameNet would define each phase as a semantic frame. Although not surprising given the situation, note that transitions between the major phases of FF are primarily communication events, all of which follow Chain-of-Command conventions.

Figure 4 captures the sub-phases of Full Force Fire Fighting (F^4), in FrameNet terms subframes of F^4. As Figure 4 suggests, during Full Force Fire Fighting, numerous actions or sub-events occur simultaneously, one of which is Search and Rescue. But for Salvage, an ongoing activity during Full Force Fire Fighting, Search and Rescue takes precedence over other activities, fire conditions permitting, and continues until the unit determines that no one else needs to (or can) be rescued.

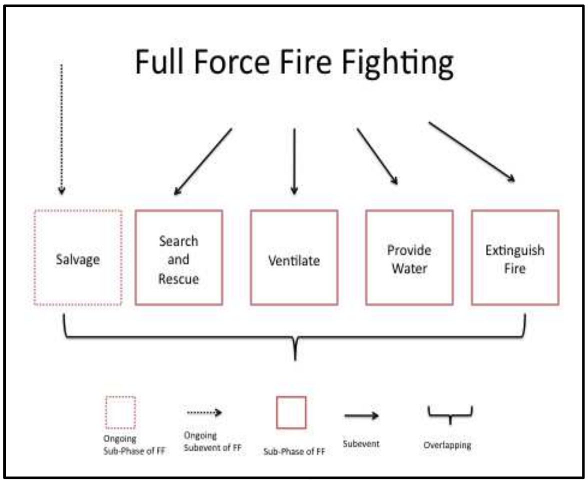


Figure 4: Full Force Fire Fighting

Fire fighters must communicate their status to the incident commander (IC), usually a battalion chief, so that the IC can manage the situation effectively. For instance, members of the unit that conduct Search and Rescue, typically medics, report their whereabouts and activity to the IC. Figure 5 depicts the subparts that constitute the more complex Search and Rescue scenario, each a separate subframe, and displays the relation Precedes holding between the relevant subframes.

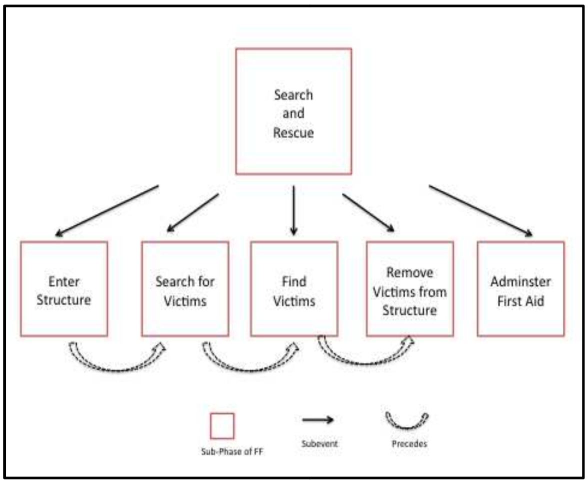


Figure 5: Search and Rescue

[6] Petruck et al. (2004) describes how FrameNet records the relations **Causative_of** and **Inchoative_of** to separate causative events, inchoative events, and statives, with each type of predicate in its own frame.

Consider (3), a possible utterance of a fire fighter conducting Search and Rescue, also updating the IC. Identifying the structured knowledge that English speakers must access to understand the entire utterance will highlight the need for **Precedes** in NLP applications. The clipped nature of (3) is more typical of speech during crisis response than of written text, thus (3) serves as a reminder that participants in Search and Rescue must exploit a wealth of shared background knowledge for effective communication.

3. Helped woman who was gasping.
   Inside the stairwell.
   Lots of smoke.
   Heading out now.

Helped evokes a scene in which someone administers first aid to a victim; and the noun phrase *woman who was gasping* describes that victim. The adverb *inside* implies that the participants are in an enclosed structure; and *the stairwell* evokes that structure. The only word that provides any hint of Fire Fighting is *smoke*. While humans can infer that the speaker of (3) entered a building, searched for, and found a gasping woman *before* administering aid, NLP systems cannot unless the system encodes information about the expected order of events in Search and Rescue. **Precedes** captures the chronological order of events in the subframes of Search and Rescue, independent of their place in the utterance's word order. The frame-to-frame information in Figure 5 allows the automatic decision about the order of events in (3). Absent **Precedes**, and other knowledge structured in the FF scenario (some of which we described), the word order of LUs in (3) could lead a system to conclude incorrectly that the main events unfolded as follows: Administer Aid, Gasp, Enter, and Exit. Because saving lives is the highest priority in Search and Rescue, Figure 5 does not show **Precedes** between Remove Victims from Structure and Administer First Aid since only the fire fighter conducting Search and Rescue can determine which order is best.

5. Conclusion

Frame Semantics is among the most useful techniques for deep semantic analysis of linguistic material, primarily text. This paper has illustrated the contribution of Frame Semantics, as instantiated in FrameNet, to research on situation awareness, highlighting the role of the relation **Precedes** for natural language understanding during crisis response. Except for Ruppenhofer et al. (2010), **Precedes** has received virtually no attention in FrameNet-related research. This work fills that gap by demonstrating the necessity of the information encoded with **Precedes** for NLP applications, advocating the inclusion of such information in systems for security applications.

Currently, FrameNet includes 82 instances of Precedes, only 4.9% of the frame-to-frame relations recorded.[7] As FrameNet continues to expand and cover more areas of English vocabulary, necessarily defining frames that characterize a greater number of complex event and state scenarios, instances of the **Precedes** relationship will increase also. Moreover, to enhance its usefulness, FrameNet must add other relations to its repertoire of frame-to-frame relations. For instance, Hasegawa et al. (2011) proposed introducing two relations new to FrameNet, i.e. **Symmetric_antonymy** (*male*/*female*) and **Asymmetric_antonymy** (*love*/*hate*) to capture different types of negation that may hold between certain LUs, enriching the FrameNet database, and facilitating its use as a resource for paraphrasing. Somewhat comparably, to facilitate inferencing work, others have suggested that FrameNet implement an entailment relationship in the database (Ovchinnikova et al. 2010).

6. Acknowledgements

DARPA Contract award #HR0011-11-C-0017 to the International Computer Science Institute in Berkeley, CA, supported part of the work presented in this paper. The authors are indebted to Charles J. Fillmore, Jim Hieronymous, Collin F. Baker, Michael Ellsworth, and Russell Lee-Goldman for helpful comments. Only the authors are responsible for the final version.

7. References

Berlin, B., D.E. Breedlove and P.E. Raven (1973). General Principles of Classification and Nomenclature in Folk Biology, *American Anthropology*, 75, pp. 214–42.

Brown, C.E., J. Kolar and B.J. Torrey, T. Trương-Quang and Volkman, P. (1976). Some general principles of biological and non-biological folk classification. *American Ethnologist* 3, pp 73–85.

Burchardt, A, M. Pennacchiotti, S. Thater and M. Pinkal. (2009). Assessing the impact of frame semantics on textual entailment. *Natural Language Engineering* 15(4), pp. 527–550.

Chang, N., S. Narayanan and M.R.L. Petruck (2002). Putting Frames in Perspective. In *Proceedings of the Nineteenth International Conference on Computational Linguistics*, Taipei: COLING, pp. 148-154.

Cruse, A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Fillmore, C.J. (1975). An alternative to checklist theories of meaning. In *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*, Berkeley: BLS, pp. 123-131.

--- (1977). The need for a frame semantics in linguistics. In H. Karlgren (Ed.) *Statistical Methods in Linguistics*. Stockholm: Skriptor, pp. 5-29.

--- (1978). On the organization of semantic information in the lexicon. In *Papers from the Parasession on the Lexicon*, Chicago: CLS, pp. 148-173.

--- (1985). Frames and the Semantics of Understanding. *Quaderni di Semantica*, 6(2), pp. 222-254.

[7] The Appendix lists examples in FrameNet, where **Precedes** holds between a frame in column 1 and its partner in column 2.

Fillmore, C.J. and Baker, C.F. (2001). Frame Semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop*, Pittsburgh: NAACL, pp. 59–64.

--- (2004). The evolution of FrameNet annotation practices. In *Proceedings of Building Lexical Resources from Semantically Annotated Corpora Workshop*. Lisbon: LREC, pp.1-8.

Fillmore, C.J., C.F. Baker, and H. Sato (2004). FrameNet as “Net”. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Paris: ELRA.Volume IV, pp. 1091-1094.

Fillmore, C.J., M.R.L. Petruck, J. Ruppenhofer, and A. Wright (2003). FrameNet in action: The case of attaching, *International Journal of Lexicography*, 16(3), pp. 297-332.

Hasegawa, Y., R. Lee-Goldman, A. Kong and A. Kimi, (2011). FrameNet as a resource for paraphrase research. *Constructions and Frames*, 3(1), pp. 104–127.

Liston, J.L. (1972). The semantic structure of body part terms in Serbo-Croatian: I, The part-whole hierarchy. *Anthropological Linguistics* 14(8), pp. 323-338

McClure, E.F. (1975). Ethno-anatomy: The structure of the domain. *Anthropological Linguistics* 17.2: 78-88.

Mehrota, S., T. Znati and C.W. Thompson (2008). Crisis Management. *IEEE Internet Computing* 12(1), pp. 14-17.

Ovchinnikova, E., N. Montazeri, T. Alexandrov, J.R. Hobbs, M.C. McCord, and R. Mulkar-Mehta (2011). Abductive Reasoning with a Large Knowledge Base for Discourse Processing. In *Proceedings the Ninth International Conference on Computational Semantics*, Oxford: ACL pp. 225-234.

Petruck, M.R.L., C.J. Fillmore, C.F. Baker, M. Ellsworth and J. Ruppenhoffer (2004). Reframing FrameNet data. In *Proceedings of The Eleventh European Association of Lexicigraphy International Congress*, Lorient: EURALEX, pp. 405-416.

Ruppenhofer, J., M. Ellsworth, M.R.L. Petruck, C.R. Johnson, and J. Scheffczyk (2010). *FrameNet II: Extended Theory and Practice* (Web Publication http://framenet.icsi.berkeley.edu).

Shen, D. and M. Lapata. (2007). Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague: ACL, pp. 12–21.

A. Appendix: Precedes holds from 1 to 2

#1	#2
Attempt	Success_or_failure
Existence	Ceasing_to_be
Get_a_job	Being_employed
Being_employed	Quitting
Hiring	Employing
Employing	Firing
Committing_crime	Criminal_investigation
Arrest	Arraignment
Birth	Death
Birth	Dying
Waking_up	Being_awake
Being_awake	Fall_asleep
Being_awake	Getting_up
Coming_to_be	Existence
Aiming	Hit_or_miss
Visiting	Visitor_departure
Invading	Conquering
Invading	Repel
Dying	Death
Confronting_problem	Resolve_problem
Ceasing_to_be	Out_of_existence
Fall_asleep	Sleep
Assemble	Meet_with
Arraignment	Trial
Notification_of_charges	Entering_of_plea
Entering_of_plea	Bail_decision
Change_of_phase	Altered_phase
Jury_deliberation	Verdict
Criminal_investigation	Criminal_process
Court_examination	Jury_deliberation
Text	Labor_product
Trial	Sentencing
Being_born	Dying
Being_born	Death
Employment_continue	Employment_end
Employment_start	Employment_continue
Visit_host_arrival	Visit_host_stay
Visit_host_stay	Visit_host_departure
Visiting_scenario_arrival	Visiting_scenario_stay
Visiting_scenario_stay	Visiting_scenario_departing
Visitor_arrival	Visiting
Event	Change_of_state_endstate
Sleep	Waking_up

Expressive speech synthesis database for emergent messages and warnings generation in critical situations

Milan Rusko, Sakhia Darjaa, Marian Trnka, Miloš Cerňak
Institute of Informatics of the Slovak Academy of Sciences
Dúbravská cesta 9, 845 07 Bratislava, Slovakia
E-mail: Milan.Rusko@savba.sk, utrrsach@savba.sk, Marian.Trnka@savba.sk, Milos.Cernak@savba.sk.

Abstract

Automatic information and warning systems can be used to inform, warn, instruct and navigate people in dangerous and critical situations, and increase the effectiveness of crisis management and rescue operations. One of the activities in the frame of the EU SF project CRISIS is called "Extremely expressive (hyper-expressive) speech synthesis for urgent warning messages generation". It is aimed at research and development enabling the possibility to design speech synthesizers with high naturalness and intelligibility in Slovak which will be capable of generating messages with various expressive loads. The synthesizer will be applicable to generate warning system messages in case of fire, flood, state security threats, etc. Early warning in relation to the above can be made thanks to fire and flood spread forecasting; modeling thereof is covered by other activities of the CRISIS project. The most important part needed for synthesizer building is the speech database. A method is proposed to create such a database. The first version of the expressive speech database is introduced and first experiments with expressive synthesizers trained with this database are discussed.

Keywords: crisis management, expressive speech synthesis, expressive speech database.

1. Introduction

Speech communication is the most common and most effective natural information transfer means used by humans. Thus, automatic information systems increasingly use interfaces allowing human speech communication between the user and the automatic system. However, for the time being artificial speech generating systems (synthesizers) are not able to generate speech with correct prosodic (more exactly, "supra-segmental", or from another viewpoint paralinguistic and extra-linguistic) features bearing additional information such as the content of emotions, the level of the message strength and urgency, its warning or reassuring tone, etc. Information systems should be able to give this important information as well, for instance in emergency situations. It is known that investigators of some aircraft accidents believe that neutral and even reassuring mode of warning messages read by synthetic voice probably contributed to incorrect evaluation of the level of danger by the pilots.

Both psychologists and engineers categorize emotions expressed in speech in different manner and into different number of so-called basic emotions (most often 4 or 6). In our work, however, we do not want to model joy, sorrow, etc.; rather, we want to create speech which would express warning and strong caution and, we want to implement reading messages or holding a reassuring dialogue.

Since almost all currently used speech synthesis methods apply large speech databases to acquire the data necessary for their development and activities (training and testing), design and subsequent creation (recording, annotation and processing) of specialised speech databases is necessary as a first step. Simultaneously, a new type of speech synthesizer is being developed, using special speech databases and modelling of features on the basis of Hidden Markov Models (HMMs) with more accurate modelling of supra-segmental features.

2. Expressive HMM-TTS for public safety

2.1. The emergency situations

In many emergency situations and collective crises of different scale the responsible management would need to make use of the information system equipped with expressive speech synthesizer capable of delivering automatic and updatable urgent messages to the needed places.

To mention just some of the critical situations, they range from everyday community emergencies, like traffic accidents, smaller fires, flooding, chemical, radioactive, or biological spills, medical emergencies, criminal activities, bomb threats, workplace violence, major power outages and many others, to disasters like large scale fires, catastrophic floods, ecological catastrophes, earthquakes, tsunami, technological catastrophes or even space catastrophes. Obviously, quantitatively different emergency situations require very different emergency management procedures.

2.2. Project CRISIS

The European Structural Funds project "Research and Development of New Information Technologies for Prediction and Solution of Critical Situations of Inhabitants" CRISIS[1] (ITMS 26240220060) is aimed at predicting and solving critical situations when the inhabitants are endangered, and the environment has to be protected.

The project is targeting:

- Creating information systems that support grid technology to provide massive computational power to solve difficult tasks during the management of critical situations
- Computer simulation of fires and its visualization

[1] http://www.crisis.sk/

- Extremely expressive (hyper-expressive) speech synthesis for urgent warning messages generation
- The area of mechatronic systems
- Nanotechnologies with special focus on sensors development
- Development of safe platforms

2.3. The role of Expressive TTS

The goal of the “Expressive speech synthesis” activity is to perform basic and applied research and to develop a system which would be capable of generating information system messages and dialogue system replies in naturally sounding speech with considerable content of paralinguistic and extra-linguistic information representing properties such as warning tone, urgency, but also soothing and reassuring speech tone. The application result would be represented by a prototype of a new speech synthesizer using large speech databases and modelling using hidden Markov models.

The synthesizer will be applicable to generate warning system messages in case of fire, flood, state security threats, etc. Early warning in relation to the above can be made thanks to fire and flood spread forecasting; modelling thereof is covered by other activities of the CRISIS project.

3. Speech resources

The speech synthesis database, which is being used in the initial phases of the hyper-expressive synthesis research and development, consists of several partial speech databases. Some parts were recorded by the “source speaker”, whose voice will be used as the source for spectral information in the final synthesized voice. The other partial databases are of different origin (dance coaching, football trainings, puppeteer performances, TV shows etc.) and serve mainly for basic phonetic research of expressive speech and experiments with prosody modeling.

In this paper we will concentrate on the set of three databases recorded by the source speaker:

- Basic database for speech synthesis in Slovak
- Read speech database for ASR training and testing
- Expressive speech database

3.1. Basic database for speech synthesis in Slovak

Neutral speech database is taken from our previous work done by Rusko at al. (2004). We briefly introduce here its content.

The database consists of about 2000 sentences pronounced by one speaker and recorded in studio conditions, divided into following categories:

- Phonetically rich sentences (the largest category, contains 1500 sentences)
- Set of words covering all Slovak diphones
- Set of sentences covering intonation phenomena
- Spontaneous speech recordings (General topic story, Application oriented story)
- Set of prompted application-oriented phrases and embedded application commands
- Numerals

The annotation consists of *two text annotation levels*: orthographic and orthoepic; *five signal annotation levels*: microsegmental information - pointers to individual pitch periods, phoneme boundaries, diphone boundaries, syllable boundaries, and whole words and phrases information; and *two suprasegmental annotation levels*: melody contour information - smoothed f0 value plus intonation phrase boundaries and accent information.

Figures 1 and 2 show an analysis (Rusko at al., 2008) of two important prosodic factors: phoneme lengths in the database and F0 values.

Figure 1: Mean segmental lengths of phonemes in the neutral speech database (in ms) and their respective standard deviations.

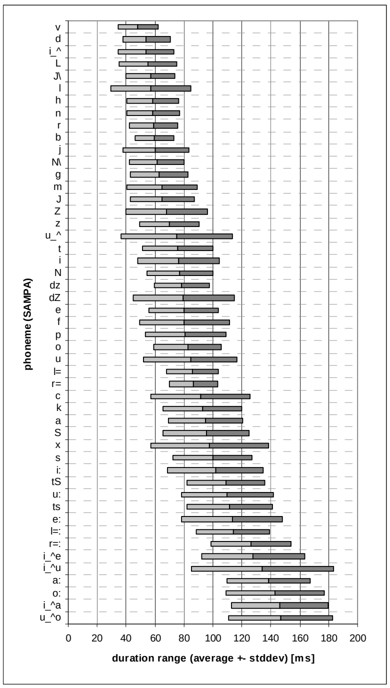
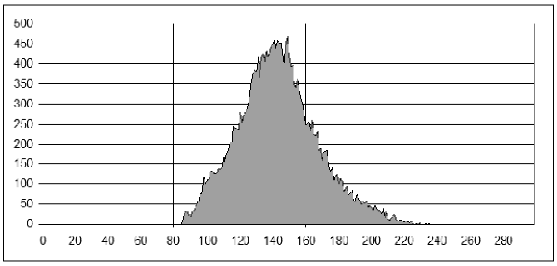


Figure 2 shows the histogram of the values of F0 at the centers of the syllables. The x axis shows fundamental frequency in Hz and y axis shows the frequency of occurrence of the given F0 value in the whole database. The mean of all F0 values at the centers of the 24968 syllables is 143.18 Hz. The mean F0 measured over speech signal in the whole database is 136 Hz.

Figure 2: Histogram of the values of F0 in the database



3.2. Read speech database for ASR training and testing

For different purposes in the area of automatic speech recognition research and development a 242 speaker speech database was created at the Institute of Informatics recently. The database consists of read sentences, mainly from the judicial domain, recorded in studio environment. One of the speakers is our source speaker. His 45 minutes of annotated speech signal are therefore also available for speech synthesis and enrich his neutral speech resources significantly.

3.3 Expressive speech database

As mentioned above, there are several speech databases being built in the frame of the Crisis project, namely:

1. spontaneous speech recordings of dance teachers of the children folk ensemble
2. spontaneous speech recordings of the football trainers
3. semi- spontaneous speech recordings of the “Law court” TV show
4. prompted recordings of short warning messages

The first three serve mainly for the basic research in phonetics and for prosody modelling experiments.

The last database is meant to be directly used in applications development. We will describe only this database in detail, because it is already finished and can be used in expressive speech synthesis.

The database was recorded in the acoustically treated studio. Rode K2 microphone was used due to its wide dynamic range.

The speech material consists of warning messages with lengths ranging from one word to four sentences. The prompted messages were uttered in three degrees of urgency. The speaker was instructed to utter the message once in a neutral manner (the first level of expressivity), then with higher imperativeness, like a serious command or directive (the second level of expressivity), and finally like an extremely urgent command or statement being declared in a situation when human lives are directly in danger (the third level of expressivity).

The recordings were then automatically annotated using forced alignment in HTK[2] with a Slovak acoustic model (Mirilovič, 2011). The annotation conventions are the same as with the Basic speech synthesis database (Rusko at al., 2004) in this version.

3.4. The short warning messages, their definition and their structure

The short warning messages are for the purpose of this study defined as short spoken messages used to warn the persons in danger. Their length typically ranges from one word to several sentences. Longer messages diminish the expressive load representing emergency in their speech signal. It is very important for the speaker to keep the voice effort high during the whole utterance and often even add accents also at the end of the utterance.

The basic parts of the Short Warning Message are the following:

1. The name and the type of the message
2. Addressing of the addressee
3. Label (e.g. “Attention!”)
4. Sender of the message
5. Time information
6. Place
7. Specification of emergency
8. Level of emergency
9. Advices and commands
10. Additional information

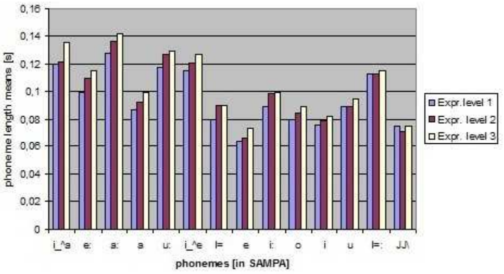
The parts of the message are optional except for the specification of emergency, which is obligatory. Following this structure a set of messages from the very short and brief ones, to more complicated and lengthy can be created for different emergency situations. This approach was used to construct textual messages which served as prompts for recording the expressive speech database.

3.5 Current state of the expressive database

The first version of the expressive database has been finished recently. 90 prompts were recorded from one speaker for each expressivity level, which represents more than 160 sentences per level.

The analysis of the acoustic-phonetic features has brought several interesting, although not unexpected, results. The changes of the phoneme lengths are different for different phonetic classes. With increasing expressive load the vowels and diphthongs are lengthened to about 105% at the second level and 110% at the third level (see Figure. 3).

Figure 3: Lengths of vowels and diphthongs.



On the other hand, consonants, mainly sibilants, are significantly shortened with increasing expressivity in our database (see Fig. 4).

Figure 4: Consonant lengths of expressive speech.

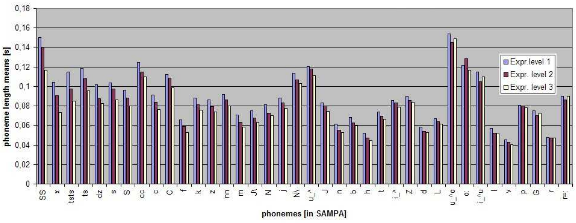
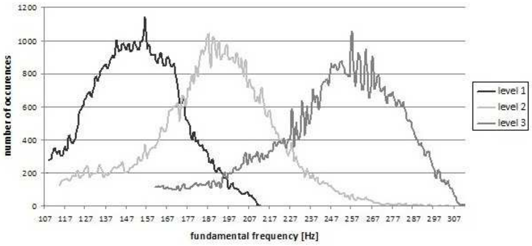


Figure 5: Pitch histograms of expressive speech.



The pitch is increasing with the expressive load too (see Fig. 5). The recording of the next part of the expressive database will continue after the results of expressive synthesis with the first version of the database are analysed in detail, evaluated and further needed extensions to the database are defined.

[2] `http://htk.eng.cam.ac.uk/`

4. Experiments with expressive speech synthesis

There are two main approaches to the creation of expressive HMM synthesizer. One is to use only the recordings with high expressive load directly for training the models (*direct synt.*). The second one is to adapt the neutral voice to a higher level of expressiveness using the expressive speech database (*adapted synt.*).

The HTS system (Zen at al., 2007) was used for experiments in speech synthesis. For the development of adapted voice a Constrained Structural Maximum A-Posteriori Linear Regression (CSMAPLR) of Nakano et al. (2006) was used as state-of-the-art HMM-TTS adaptation technique. We have tried both approaches and for both the results are very promising. In spite of the relatively low number of utterances recorded, we were able to train a functional synthesizer with a very high level of expressiveness. According to our informal listening tests the synthesized speech keeps the voice-quality, rhythmical and pitch features from the source recordings very well.

Figure 6. illustrates the results of the expressive speech synthesis. It presents the spectrogram and oscillogram of one utterance “Hrozí únik plynu!” (*Imminent danger of gas-escape!*) at the third level of expressivity produced by a human speaker (left), *adapted synthesis* (middle) and *direct synthesis* (right) methods.

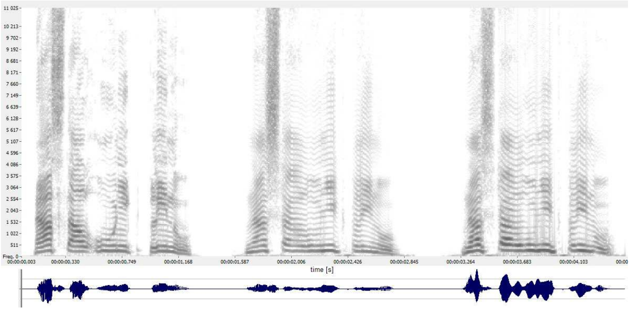


Figure 6. The spectrogram and oscillogram of one read and two synthesized utterances “Hrozí únik plynu!” (*Imminent danger of gas-escape!*) at the highest level of expressivity.

While the *direct synthesis* method keeps the timbre and intonation very similar to that of the original speaker’s expressive speech, the overall quality of the *adapt synthesis* method seems to be a bit higher with less disturbing noise in the signal. This is probably caused by much bigger amount of training data available for neutral speech, which were not used in the previous method. On the other hand the expressive load is a bit lower for the *adapted synthesis*.

4. Conclusion

We introduced a collection of speech resources that can be used for the development of hyper-expressive speech synthesis in Slovak, aimed to be used in the public safety domain. The paper focuses on the expressive speech database consisting of the annotated recordings of 90 prompted short warning messages (160 sentences per level) uttered by one male speaker at three levels of urgency. The first level is neutral speech and serves mainly for comparison with the other two levels. The second level represents assertive warnings or commands, and the third level produces extremely intense and urgent messages when lives are endangered.

Preliminary experiments with HMM speech synthesis were performed which imply that the proposed method of expressive speech database development is suitable for gathering a good quality expressive and hyper-expressive speech database for the design of speech synthesizers for emergency situations.

It is worth mentioning the following three issues:

1. Data recording. The hyper-expressive messages were often uttered with voice quality and loudness close to shouting. Sometimes there was also panic noticeable in the voice. This shows again that a lot of information is carried by the suprasegmental phenomena and it is very difficult for the actor to set his voice correctly for one purpose – uttering warning messages.

2. Data analysis. The position of formants seems to be relatively stable but the F0 mean is increasing and the strong change of spectral slope is evident from the increasing content of higher frequency components. These phenomena together with typical pitch contours and rhythmical patterns are to be modeled very precisely and so even more of the domain dependent expressive data will be needed in future.

3. Building of expressive TTS. Two expressive HMM synthesizers were compared using our informal listening tests. The direct synthesis from the most expressive data is possible for a constrained domain usage, and it sounds authentically, but it suffers from missing some context dependent phonemes in the database due to its limited volume. We think that the adapted synthesis based on a slightly modified general adaptation approach produces results of higher overall quality of the synthesized hyper-expressive voice at the expense of a bit lower expressive load of the produced speech.

5. Acknowledgements

This publication is the result of the project implementation: CRISIS, ITMS 26240220064 supported by the Research & Development Operational Programme funded by the ERDF.

6. References

Rusko, M., Trnka, M., Darjaa, S. and Cernak, M. (2004) *Slovak Speech Database for Experiments and Application Building in Unit Selection Speech Synthesis*. In P. Sojka, I. Kopecek, and K. Pala, editors, Proceedings of TSD 2004. Springer, Brno

Rusko M., Trnka M., Darjaa S., Kováč R., Hamar J. (2008): *Modelling acoustic parameters of prosody for read and acted-speech synthesis.* In: Proceedings of Acoustics '08 Paris, France, pp. 1273-1278.

Darjaa, S., Trnka, M., Cernak, M., Rusko, M., Sabo, R., and Hluchy, L. (2011). *HMM speech synthesizer in Slovak*. In GCCP 2011 : 7th International Workshop on Grid Computing for Complex Problems. Bratislava : Institute of Informatics SAS, 2011, p. 212-221.

H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda (2007), The HMM-based speech synthesis system version 2.0, Proc. of ISCA SSW6, Bonn, Germany.

Nakano, Y., Makoto, T., Yamagishi, J., Kobayashi, T., (2006), Constrained Structural Maximum A Posteriori Linear Regression for Average-Voice-Based Speech Synthesis, Proc. of ICSLP’06.

Mirilovič, M., (2011), Akustické modelovanie v ARR. In JUHÁR, Jozef. Rečové technológie v telekomunikačných a informačných systémoch. - Košice : EQUILIBRIA, s.r.o.

Language Resources in a Public Security Application with Text Understanding Competence. A Case Study: POLINT-112-SMS

Zygmunt Vetulani
Adam Mickiewicz University, Poznań, Poland
E-mail: vetulani@amu.edu.pl

Abstract

The aim of this paper is to show the importance of language resources in the development of complex, public security oriented applications with natural language understanding components as essential parts of the system. We present a case study of a mature project in the public security sector. This case study aims at giving an idea of the spectrum of needs and problems, without pretention to exhaust the topic. As it is typical for public security oriented projects, besides usual problems due to the gaps in available language data (resources), designers and developers of the presented system needed to deal with sensitive data necessary for efficient language modeling. To make the paper self-contained, we start with a compact presentation of the POLINT-112-SMS system. Then we present the language resources we used.

Keywords: language resources, public security, mass event

1. Introduction

The idea of considering the language industry as an autonomous sector of economical life and language resources as its main branch is commonly attributed to Antonio Zampolli who launched it in a visionary way in the 1980s. The success of LREC conferences confirms the validity of this vision. Within the progress of the idea of the Global Information Society, language technologies quasi penetrated all fields. The aim of this paper is to show the importance of language resources in the development of complex information processing systems in the domain of public security, where in many applications natural language understanding components play an essential role. The demonstration is made in form of a case study of an mature project in the public security sector. The nature of the application makes that besides the usual problems, due to the gaps in existing and available language resources, the system designers need to deal with sensitive data necessary for efficient language modeling. This case study intends to give an idea of the spectrum of needs and problems, without however pretention for the study to be exhaustive. To make the paper self-contained, we start it with a compact presentation of the POLINT-112-SMS system including its architecture. Then we focus on particular resources used in the system design and implementation.

2. The POLINT-112-SMS system

2.1. Logical model of the system

The POLINT-112-SMS[1] system is designed in order to assist competent public services (typically Police in Poland) in the tasks of security protection at large scale mass events[2]. Its main role is assisting the monitoring process of mass events and real-time identification of processes in the crowd of fans in order to discover potentially dangerous situations with a high degeneration risk (early prevention). The efficient monitoring is realized through direct observation of the event by the trained staff (informers) being in contact with the Emergency Center. It is recommended, for the efficiency and security of the personnel, to organize the communication channel in a possibly hidden way (to avoid disclosing the informers and provoking fans). This postulate is satisfied through the use of the SMS-like, text-based message transmission mode. The second important reason for using the text mode is that operations are typically performed in a very noisy environment.

These two reasons make that POLINT-112-SMS is an example of a system where text-based communication is the primary communication model. The application has the following characteristics.

- The system subscribers[3] are typically
  - informers (e.g. Police staff) supposed to transmit information about people and events to the Emergency Center,
  - operator (decision making head of the emergency service or his/her assistant),
  - analysts whose role is to process information already preprocessed by the POLINT-112-SMS system.
- The subscribers address information or information requests to the system behaving as an expert which helps in the decision making process (POLINT-112-SMS doesn't take any decision autonomously, these are reserved for humans).

[1] The reader may find many information on the POLINT-112-SMS project in (Vetulani et al., 2010; in Polish).

[2] A typical example of such mass event is a football match with a large number of fans (several thousand). This is the setting chosen for our studies.

[3] We identify the following three categories of the system users: Informer, Analyst, Operator.

- Dialog between the subscribers and the system is performed through short text messages (SMS) and without human aid.
- The system identifies facts (events), process and stores them for further re-use.
- The system may transfer knowledge to authorized subscribers through question answering (text).
- The system visualizes events and the current situation at the stadium to help human decision taking.

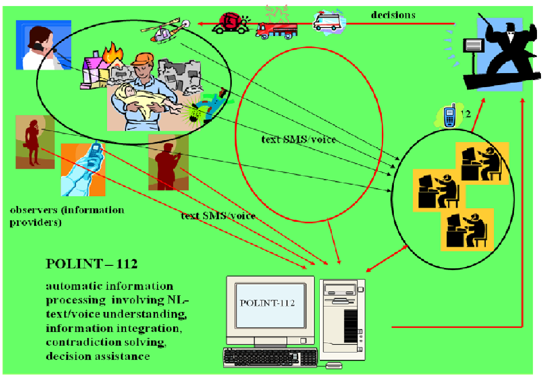

decisions
observers (information providers)
text SMS/voice
text SMS/voice
112
POLINT – 112
automatic information processing involving NL-text/voice understanding, information integration, contradiction solving, decision assistance
POLINT-112

Fig. 1. A use case scheme for POLINT-112-SMS (example)

The main feature of the system is its possibility of interaction between subscribers and the system in the dialogue mode (close to natural, spontaneous dialogue). The model is generic for a large class of applications, in particular in the security maintenance, crisis management at natural or technological disasters etc.

The following are typical requirement/needs which actually are satisfied by POLINT-112-SMS, or may be implemented in its extensions (provided an additional R&D effort):

- Possibility to dialogue between the user and an emergency service (voice[4], SMS, Instant Messaging systems...).
- Possibility of automatic monitoring of the human-to-human emergency dialogue in order to support the human operator and replace him/her in some routine operations as gathering facts and collecting personal data e.g. through reminding him/her the procedures that he/her has to apply.
- Searching for complementary (confirming) information (systems ask questions to other informers if it finds it necessary).
- The system looks for correlations between findings made by different informers independently and try to formulate and to verify new hypotheses about threats; automatic (and autonomous) discovery of threats.
- The human analyst may formulate requirements concerning the knowledge that is to be collected and force the system to search for that knowledge.
- The systems may share its knowledge with the authorized users (on request or automatically).
- The system may process (in real time) a huge amount of users' messages autonomously.

Let us notice that a system with all above functionalities (in particular with the voice recognition) will satisfy all typical requirements and expectations one can address with respect to the traditional "112" service with however the absence of some of its well known drawbacks (as small processing capacity and frequent communication bottlenecks).

2.2. System architecture

Information stored in the POLINT-112-SMS system is represented using three data structure types: Message, Event, Situation. Each of these data structure deals with different kinds of information entering to the system at different stages.

The POLINT-112-SMS system architecture (simplified) is composed of 6 main modules (and some sub-modules).

1) The SMS Gate - enables SMS-based communication between human agents and the system. It is composed of two sub-modules: for sending messages, and for receiving them. The SMS Gate communicates directly with the NLP Module.

2) The NLP Module - for message text analysis (lexical, syntactic and semantic) and generation. Its main function is surface understanding of users' messages. The NLP Module communicates directly with the SMS Gate and with the DMM. The NLP Module (supported by the DMM) focuses on transforming single sentences into data structures without storing and processing these structures.

3) Dialogue Maintenance Module (DMM) - is responsible for the deep understanding and dialogue with the user. It performs anaphora analysis and disambiguation. DMM communicates directly with the NLP and SAM modules.

4) Situation Analysis Module (SAM) - the brain of the system - reasons about the structures without communicating directly with the informers. It is responsible for information processing and knowledge management, contradiction solving (reasoning, information seeking). The SAM controls a number of subordinate modules (responsible for supplying the ontology (PolNet), storing general knowledge, reasoning about space and time, event recognition, identification of objects etc.).

5) World Knowledge Module stores permanent knowledge about the project universe. It communicates with the NLP Module and with SAM. It controls general information about the world concepts which is first of all contained in a language inspired ontology (based on PolNet).

6) The Visualization and Control Module displays on the screen the pictorial symbolic representation of current information and permits the user to make and store snapshots for further processing. It communicates directly with SAM. It serves also as system administrator's gate.

[4] Voice communication requires a high quality STT system, nonexistent for Polish now (at the required level of excellence level), but available for some other languages.

Informer: Kowal nosi bluzę i czarne spodnie. Kowal i Wolski są na murawie. Kowal i Wolski należą do bojówki. *(Kowal is wearing a blouse and black trousers. Kowal and Wolski are on the football field. Kowal and Wolski are members of a hooligan squad)*

System: 'Zrozumiałem' *(Understood)*

Informer: Osoba, która nosi czarne spodnie, jest niebezpieczna. Kowal i Wolski atakują policjantów kamieniami. *(The person who wears black trousers is dangerous. Kowal and Wolski are attacking policemen with stones)*

System: 'Zrozumiałem' *(Understood)*

Analyst: Kto atakuje policjantów? *(Who is attacking the policemen?)*

System: 'Kowal. Wolski. Brak dalszych odpowiedzi!' *(Kowal. Wolski. No more answers!)*

Analyst: Kto jest niebezpieczny? *(Who is dangerous?)*

System: 'Kowal. Wolski. Brak dalszych odpowiedzi!' *(Kowal. Wolski. No more answers!)*

Analyst: Czy istnieje niebezpieczna osoba, która nie została złapana? *(Is there a dangerous person who has not been captured?)*

System: 'Tak, kilka, w szczególności istnieje: Kowal !' *(Yes, a few persons, in particular Kowal!)*

Informer: Kowal ucieka. *(Kowal is escaping)*

System: 'Zrozumiałem' *('Understood')*

Informer: Kowal został złapany. *(Kowal has been captured)*

Fig. 2. A simple dialogue with a POLINT-112-SMS prototype[5]

3. Language resources for language application development

In the project we have used and developed generic language resources like dictionaries, grammars and ontologies. For development purposes we have also used our own resources developed within the project and primarily for the project. Some of them are now public (or are going to be public).

3.1. Publicly accessible resources

Among the resources which were publicly available at the project start date (2006) we used the following ones:

1) Morphological electronic dictionary POLEX (and its derivatives)
2) IPI PAN Corpus
3) Universal Dictionary of Polish (UDJP, Uniwersalny Słownik Języka Polskiego)
4) Internet dictionary SJP.PL
5) Generative Dictionary of Polish Verbs (Generatywny Słownik Czasowników Polskich)

3.1.1. POLEX

The electronic dictionary POLEX[6] (distibuted[7] through the Evaluations and Language resources Distribution Agency (ELDA), www.elda.org) was the main source of the morphologic information used in the project. The generic version of the dictionary was developed in the Dept. of Computer Linguistics of the Adam Mickiewicz University during 1993-1996 within a Polish Government funded project[8]. The dictionary contains over 100,000 entries encoded in an human-and-machine readable format, the same for all parts of speech (Vetulani et al., 1998a,b). POLEX contains the main Polish vocabulary of general interest. Entries have the following format:

```
BASIC_FORM +
          LIST_OF_STEMS +
                    PARADIGMATIC_CODE +
                              STEMS_DISTRIBUTION
```

For example, the dictionary items for frajerI and frajerII are as follows:

frajer; frajer,frajerz; N110; 1:1-5,9-13; 2:6-8,14
frajer; frajer,frajerz; N110; 1:1-5,8-14; 2:6-7

It includes the list of words of the Dictionary of Polish Language PWN (Szymczak (ed.), 1981) which entitles us to claim that it covers the lexical kernel of the Polish language (with over 42,000 nouns and 12,000 verbs). The existing gaps, in particular those essential for our application were completed within the project using our own tools (in particular the UAM Text Tools)[9].

3.1.2. IPI PAN Corpus

The IPI PAN Corpus was the first large scale project corpus pretending to the role of reference corpus for contemporary Polish. For this research we had at our disposal a 80 million words fragment (non-annotated) of the IPI PAN Corpus[10] whose size, at the project start date (2006), was about 200 million traditional words. Nowadays, its successor contains some 1,5 billion text words and it is known as the National Corpus of Polish (NKJP); cf. (Przepiórkowski et al., 2011).

3.1.3. UDJP

We used also the Universal Dictionary of Polish by Dubisz (2006). This is a monolingual (human readable only) dictionary in the tradition of more classical, earlier dictionaries by Doroszewski (1958) and Szymczak (1995). It provides inflectional and semantic information, as well as usage examples for some 100,000 entries. However it cannot be used directly within the computer software because of its format and lack of syntactic information (in particular for verbs).

3.1.4. SJP.PL

SJP.PL is a free software (under PGL and CCShare Alike licenses) developed by internauts for orthography check. It contains some 180,000 entries composed of lemmas with derived forms.

[5] The example from (Vetulani et al., 2008).

[6] (Vetulani, Z., 2000).

[7] A symbolic fee is applied for research purposes license. A simple P-O-S tagger is included in the package (in PROLOG).

[8] "POLEX - Polski Leksykon Morfologiczny" (1993-1996) headed by Zygmunt Vetulani.

[9] By T. Obrębski, cf. (Obrębski & Stolarski, 2006).

[10] http://korpus.pl/index.php?lang=en; access Mars 18, 2012.

3.1.5. GSL

The Generative Syntactic Lexicon of Polish Verbs (GSL) is the only publicly available dictionary (paper edition only) with detailed information concerning the Polish verb valency. This lexicon characterizes the syntactic and semantic connectivity of over 10,000 verbs forming the core of the Polish verbal system. The GSL entries are organized as follows:

a) entry identifier (verb in infinitive) (*lemma*)

b) optional meaning description (informal) when necessary for meaning differentiation,

c) formula (or formulae) (called by Polański *sentential scheme*) showing the syntactic structure and syntactic requirements of the verb with respect to obligatory and optional arguments,

d) specification of semantic requirements (*semantic class*) of the verb with respect to the obligatory and facultative arguments (*syntactic_frame_slot*),

e) examples of use (natural language).

3.2. The project's own resources

For the purposes of the project we had to develop the project's own resources among which were:

1) corpora,
2) dictionaries,
3) grammars,
4) ontologies.

We briefly describe them below.

3.2.1. Corpora

The project development was based on the iterative model which involves experiment-implementation-testing cycles (cf. (Vetulani, Z. 1989) and (Vetulani, Z. & Marciniak 2000)). Application of this experiment based methodology results with corpora which, is some cases, may be reused for similar applications. In the project we have collected and explored various kind corpora: text and dialogue written corpora, SMS corpora, recorded telephone conversations.

3.2.1.1. Text corpora

- **Printed texts (electronic source)**

Text corpora are typically used in order to build the language models necessary for application design. One of the most important characteristics of a language model is its lexical structure. In order to correctly construct the application's dictionary we had to set up a corpus composed of texts which:

- are from the security field,
- represent a variety of text types (book chapters, research papers),
- are representative for public security (with special attention to terrorism).

We used for this[11]:

- two books[12] (2 authors),
- research papers (23 authors)[13],
- legislative texts on terrorism.

Legislative texts were selected using the Lex Omega program. Documents represent the state-of-law in 2006 reflected in the legislation from the period 1970-2006. Documents were selected by a search procedure controlled by keywords like: "terroryzm" (*terrorism*), "terrorysta"(terrorist, n.), "terrorystyczny" (*terrorist, adj*.) and "finansowanie terroryzmu" (*terrorism funding*) and "przestępczość zorganizowana" (*organized crime*). The documents belong to 3 groups: laws (27), international conventions (41), lower level acts (12). This small corpus (some 1,000,000 tokens) permitted acquisition of terminology rarely observed or absent in general corpora (IPI PAN), especially in the area of operational police activity and terrorism.

- **SMS dialogue corpora[14]**

Dialogue is the most common form of everyday language usage but real (spontaneous) dialogue corpora are rare. Dialogue appear frequently in written texts but, in most cases, in form of fictional literature utterances with little value for natural dialogue modeling (although, in most cases, authors intend to make them „natural”). The lack of recorded, natural dialogue corpora poses a problem for systematic studies within an empirically oriented methodology.

The situation with SMS data is even worse because of often very personal or often confidential content. On the other hand, due to the written-text-based nature of the SMS communication, there exist good technical solutions for SMS corpora collection. An original method of building a corpus of SMS messages was described by Fairon and Paumier in (Fairon, C., Paumier, S. 2006). This method inspired us in the acquisition of an SMS corpus within the POLINT-112-SMS project.

First, it was necessary to spread information about SMS collection for strictly research purposes among the potential providers of SMS messages. Three ways of the SMSs provision were proposed. The first method used a special internet page which was created as an aid to transmit data together with corresponding metadata concerning the message sender and his or her telephone number. From this telephone number the SMS author was supposed to send the data to the number displayed after completion of the form. The terminal GSM collected the data and forwarded them to the data base. The second method used internet as well and consisted in hand-made copying of real messages to an appropriate form. The weak side of this method was that it is open to various kinds of re-writing errors from one hand, and to the normative temptation to correct spontaneous error and to "improve" the text with respect to the original. Only a few

[11] Corpus collected Wojciech Filipkowski.

[12] Zb. Rau (2002) and W. Filipkowski (2004)

[13] From (Januszka & Gembara, 2005).

[14] Collected and described by J. Walkowska, PhD Thesis (Walkowska 2012). In this and the next sections we largely quote after Walkowska.

people used this method (in order to avoid paying for forwarded messages). The third method applied in the project consisted in a direct copying of messages – both incoming and outgoing - from telephone to computer in form of text files. (This technique is possible with some modern telephones only). Having collected the send and received messages into dialogues we obtained a corpus of some 1,700 SMS composed of 1,843 sentences (24,599 words). This corpus was very useful at the system implementation stage, as it permitted us to discover various new phenomena, important for correct functioning of beta prototypes.

- **SMS experimental corpora**

The iterative methodology of the NL-competent systems design presupposes that the early phase modeling is based on empirical study of a corpus obtained in experiments involving participation of a *Wizard of Oz.* Then, at successive iterations of the system design loops further observations are made, with – where appropriate – the Wizard of Oz participation as well.

The initial experiment was a kind of Role Playing Game with the game master, system and informers. Initially, the roles of the game master and the system were assumed by the members of the project and those of informers – the PPBW[15] experts. The role of the game master was to define and to present to the users the current situation. The necessary information was presented in form of illustrations representing various types of events. In some cases adding some text to the pictures was necessary. In such cases we took care to limit the language usage to the strict minimum (just to make possible scene understanding (E.g. Jan Kowalski, known to the Police) in order to avoid influencing the human (language) reactions. The role of the master of the game role was to instruct the users, answer their questions and solve problem connected with the experiment. The role of informers was to inform the system of all that will seems relevant to the problem (public security). Persons playing the role of the system were responsible for attesting reception of information, and for requesting additional information, necessary if some former information was not clear enough. The corpus contained 1,374 sentences. The informers were also entitled to ask question about the observed persons. One session took some 30 min. The corpus obtained in such a way was composed of dialogues between the system and the informers (exchange between the informers and the master of the game were omitted). The corpus permitted the acquisition of the vocabulary specific for police informers. A list of new words, as well as the description of syntactic and semantic requirements were created. Among other interesting and important phenomena were information taken into account at the first iteration of the system design procedure, e.g. information concerning anaphora, the thematic structure and the answer time. At the later stage, similar experiments were arranged where the partially implemented system was assisted by the wizard of Oz (cf. Vetulani, Z. in (2009)) in order to make the dialogue fluent. These experiments were necessary to confirm the correctness of the dialogue model, to detect its gaps and complete them as well to improve the language coverage. The typical intervention of the *wizard* consisted in:

- informing the user that the system didn't understand him (e.g. message containing unknown words or structures)
- replacing the user's message by its paraphrase

Five experiments provided small corpora (respectively 221, 285, 272, 222 and 198 messages) which permitted to make interesting observations concerning phenomena specific for the SMS communication. Interesting observations were made for the typology of the most common differences between the standard Polish and its SMS variant. Among the most frequent problems we noticed:

- frequent orthographic errors (due to the non-usage of orthographic correctors)
- many typing errors
- word errors due to incorrect usage of T9
- ad hoc abbreviations (beginnings of words)
- abuse of abbreviations and rare acronyms
- systematic omission of vowels
- CamelCase (concatenation of words with capitalized initials)
- emotikons
- ad hoc onomatopeas such as "hehehe"

The detailed analysis of this corpus permitted us to observe and describe interesting phenomena concerning the structure of multiagent and multithematic SMS dialogue. Good understanding of these phenomena are crucial in modeling multi-agent communication (see e.g. Walkowska 2009) and also (Vetulani et al. 2010)).

3.2.1.2. Speech corpora

Voice-based communication is substantially different from the written-text-based one, i.e., the natural voice-based speech is not necessarily structured in the same way as the natural text-based one. Therefore, in order to design a continuous speech-input-based understanding system it may not be enough to take an existing natural text-input-based understanding system and interface it with a STT interface. On the other hand, analysis of spoken dialogue corpora (of the recorded natural dialogues) may be an important source of information useful to design systems with text interface.

Within the POLINT-112-SMS project we used recorded dialogues from the "997 Polish public emergency service[16] (maintained by the Police) in order to acquire the

[15] PPBW stands for "Polish Platform for Homeland Security".

[16] Recorded telephone 997 police emergency service (also accessible through 112 service). The data were made accessible by the Police Headquarters of the City of Poznań (Wojewódzka Komenda Policji w Poznaniu) for strictly research purposes.

lexical data and necessary grammatical information. This was a case of sensitive data very useful for system design where special negotiations with the data owner were necessary.

As a result we got access to 23 469 recorded telephone connections registered between October 10, 2006 and November 30, 2006. Finally, 1818 connections corresponding to 24 h. of continuous speech were transcribed into text and tagged with a set of 16 tags indicating elements relevant for the efficiency of both STT (VTT) and understanding[17]. Among the tagged elements were mistakes, regionalisms, non-verbal elements, individual speaker's marks etc. Transcription was made with the help of the program ELAN + Eudico Linguistic Annotator V.2.6.3. licensed by Max Planck Inst. for Psycholinguistics and by our own software ETRAN[18] to re-edition and further processing. The processed material was composed of conversation samples between Police clients and Police, between Police agents and between Police and other services. The transcribed dialogues (58,935 tokens) contributed to the design the language coverage of POLINT-112-SMS.

3.2.2. Dictionaries

3.2.2.1. Further extension of the electronic morphological dictionary POLEX

Within the project we extended the morphological dictionary POLEX (Vetulani, 2000) distributed by ELDA[19] (cf. 3.1., above) to the new version, called POLEX/PMDB[20]**.** This extension was done on the basis of SJP.PL accessible within the Creative Commons Share Alike license. With only little human aid, it was possible to add to the generic POLEX almost 100,000 new entries:

nouns 41701
    common nouns 18123
    proper nouns 23578
verbs 9335
adjectives 19692
adjective passive participles 9121
adjective active participles 4941
adverbial past participles (*imiesłowy przysłówkowe uprzednie*) 7189
adverbial present participles (*imiesłowy przysłówkowe współczesne*) 4955
together 96943

Its manual valuation is not finished yet. In a sample of 464 entries (0,5% of the total) the proportion of errors did not exceed 2,5% (in most cases one incorrectly inflected form). This work was done using our own toolkit[21].

3.2.2.2. The dictionary of verbo-nominal collocations and its extension[22]

It is well known that the role of semantic predicate (Npred) may be played by an abstract noun (Gross, M. 1981; Vetulani, G. 2000). In that case it constitutes the logical center of an elementary sentence whereas the accompanying verb (called support verb (Vsup); pol. *czasownik podporowy*) doesn't have a predicative role. The typical formal model of sentences involving support verbs are as follows (N0 + Vsup + (MOD) + Npred + (Prep) + N1 + (Pred) + N2...). The support verb in Polish plays an important auxiliary role in the interpretation of the predicate. This is due to its possible metaphorical and emotional aspect. The support verb may bring information on the addressed language register, aspect (perfective / imperfective) or action mood (inchoative, terminative, progressive etc.). This means that the "predicatively empty" support verb still contribute to the sense of the collocation and to its disambiguation. It may therefore be used as classification basis for predicate nouns (Npred).

There are several reasons to set up a dictionary of verb-noun collocations. One is the frequent impossibility to translate such collocations word by word. Another one is that sometimes such collocation may form compound verbs which do not have lexicalization (as a single, one-morpheme word) in the given language, while it does have one word synonyms in another language.

The formal description of collocations is based on elementary sentences of the type subject+verb+complement(s) represented as predicate-argument(s) logical forms. The collocation dictionary, entries are composed of collocations together with arguments in order to characterize one particular meaning. The entry contains (directly or indirectly) information on:

- the grammatical aspect (perfective/imperfective) of the support verb, grammatical gender and number of the Npred,
- the grammatical case of the Npred and all its arguments (the Polish language is highly inflected),
- the way these arguments are linked to the Npred (with or without prepositions).

When the collocation requires a modifier (usually an adjective), the tag MOD is used. The formal pattern of an entry is therefore as follows: verb-noun collocation / VSup (case of Npred) / (MOD) / N1Prep(case) /

[17] Rules for transcription and tagging were defined by Z. Vetulani and A. Dąbrowski, the encoding algorithm in form of written encoding instruction was written by Agnieszka Vetulani (following instruction by Z. Vetulani). The encoding and technical staff (A. Ćwiąkała, Sz. Drgas, J. Nowak, A. Kuczma, A. Vetulani i W. Wojciechowska) were informed about confidential character of the data and deposed a written non-disclosure agreement.

[18] Mariusz Tański and Krzysztof Sielski under supervision of Zygmunta Vetulani; ETRAN permits browsing and transcribing of recorded speech.

[19] Evaluations and Language resources Distribution Agency, www.elda.org; ELDA charges a symbolic fee if the resources are used for non-commercial purposes. A simple POS tagger written in PROLOG is a part of the package.

[20] PMDB stands for Polish Morphological Data Base.

[21] UAM Text Tools (Obrębski & Stolarski, 2006).

[22] This chapter contains large citations from (Vetulani, G., 2012).

N2Prep(case)/...

The present dictionary of verb-noun collocation was built in two steps. The first one was composed of some 5,400 collocations extracted from the existing monolingual dictionaries of Polish and described in the above format. The extension on the basis of a corpus exploration[23] contains some 16,000 collocations, at the total extension cost of 8 man-months of labor.

To obtain this extension the use of specialized tools[24] was necessary. The procedure of extending the dictionary was described in detail by G. Vetulani (2012).

3.2. Extension of POLINT grammars

The POLINT grammar constitutes the kernel of the POLINT-112-SMS system and defines its linguistic competence. It is being developed since the 1980s and its predecessors served in a number of NLP-based system prototypes. The POLINT grammar is implemented in the form of rules similar to the DCG rules (Perreira & Warren 1980) which may be used to encode structural grammatical information[25]. The POLINT grammar is composed of translations of context free like rules into Prolog with, additionally, parameters (in form of general terms) in both terminal and non-terminal symbols. Additionally, we allow tests, actions and special procedures calls in the body of a grammar rule. Terms used as parameters are a natural way to express constraints (as e.g. agreement and government – frequent and systematically used in Polish) and/or sub-categorization (cf. Colmerauer (1978) in his paper on metamorphosis grammars)[26]. Procedures may be used to control parsing execution by heuristics which may speed-up the parsing on the basis of additional information obtained at the pre-analysis stage (in linear time)[27]. Interpretation of grammar rules may trigger execution of special actions as e.g. consisting in execution of semantic calculi at the parsing time. (Describing both syntax and semantics by the same rules appears very useful while grammar development/modification but generates an expensive time overhead when backtracking. The simple way to deal with this problem is the two-run analysis. In the first run (parsing), the semantic actions are not activated. At the second run (deterministic, because controlled by the parse tree calculated at the first run) the semantic actions are activated and produce the semantic value of the whole processed unit.)[28]

The linguistic coverage of a NL understanding system is defined by the dictionary and the full grammar. By *full* grammar we mean syntactic rules together with corresponding semantic interpretation procedures. The former prototypes of POLINT systems were mainly used for demonstration and teaching purposes and their coverage was lexically restricted (limited to a few thousand word forms). POLINT-112-SMS has access to a large dictionary of word forms (1,400,000) in the format POLINT, generated from POLEX (cf. Vetulani in (2004)).

The initial competence of the generic POLINT grammar (at the project start) permitted us to analyze and to understand affirmative sentences and questions from a large (open) subsystem of Polish. In particular the following types of questions were allowed:

– whether-questions (*Czy*+affirmative_sentence?)

– questions about arguments (*Who/Kto...?,What/ Co...?, With whom/ Z kim...?,* etc.)

– questions about place (*Where is…/Gdzie znajduje sie...?*)

– questions about time (When/*Kiedy...?*)

– questions about existence

– questions about name (*What is the name of…/Jak nazywa się...?*)

– questions about ontological type, position in a hierarchy (*Whom is…/Kim jest...?*)

– questions about complement (*Whose brother is …/Czyim bratem jest...?*)

At the predicate-argument level of the sentence the surface word order is free.

POLINT identify correctly a wide class of noun phrases such as:

– proper names (*Poznań*),

– compound proper names (*the city of Pozań/miasto Poznań*),

– compound phrases (*organizator i sponsor meczu*),

– common names (*miasto*),

– pronouns (*ktoś, ja, on*),

– genitive phrases – possessive (*Peter's picture/obraz Piotra*),

– genitive phrases – complement of noun (*the date of match/data meczu*).

Noun phrases may also be complemented by relative clauses (also embedded), participle phrases, adjectives etc.

[23] We used a part of the non-annotated version of the IPI PAN Corpus (Przepiórkowski, 2004) of 80 milion words.

[24] Created solely to achieve this goal; cf. papers co-authored by G. Vetulani, Z. Vetulani and T. Obrębski (2006, 2007, 2008).

[25] Pure DCG specification would be executable but extremely inefficient.

[26] Which allows easy and natural refinement of the grammar.

[27] Pre-analysis – first described in "Lexical preanalysis in a DCG parser of POLISH, in: Eberhard Klein, Françoise Pouradier Duteil, Karl Heinz Wagner (ed.), Betriebslinguistik und Linguistikbetrieb. Akten des 24 Linguistischen Kolloquiums, Bremen 1989, (Linguistisches Arbeiten 260/261), Max Niemeyer Verlag, Tübingen, 1991, p. 389 - 395." – appears very efficient. A well constructed heuristic permits – on the basis of morphosyntactic information (morphology and valency) combined with the switch based technique to execute the prolog encoded grammar - to reduce the complexity down to linear in an important number of cases. (For th switch technique, you may consult a note by Vetulani in Logic Programming Newsletter (1994), p. 10.).

[28] Described in detail in various publications, e.g. (Vetulani, 2010).

Predicates represented by verb groups may accept one, two or three arguments (referential or locative)[29]. Predicate groups may be:

1) personal forms of verbs,

2) constructions with the auxiliary *is* (*jest*) with a noun in the instrumental case or with an adjective group,

3) constructions with a support verb

The verb groups may contain adverbial modifiers.

This generic model of POLINT grammar appeared not sufficient for the intended applications, and we decided to extend it according to the results of application oriented field experiments with the beta prototype. We had to extend both the application dictionary and the grammar. Some examples provided below represent categories frequent in the experiment generated corpus:

- enumerative nominal groups (alternation of various conjunctions, here "z" and "i"),
- some free-word-order phenomena (within the compound predicate); free adverb,
- compound predicates (multipredicative constructions),
- highly elliptical constructions.

3.2.4. Ontologies

3.2.4.1. PolNet - Polish Wordnet

Within the project POLINT-112-SMS we made extensive use of ontology. This was the reason for developing PolNet – a lexical data base system of the type of Princeton WordNet built from scratch for Polish nouns (following the so called "merge model" methodology). The PolNet project started as a part of the POLINT-112-SMS project in 2006. DEBVisDic (Pala et al., 2007) has been used for synsets generation and for editing hyponymy /hyperonymy relations which are the basic relations organizing the noun part of PolNet. The project started with creation of synsets for nouns in an incremental way i.e. starting with general and frequently used vocabulary. More precisely, we selected the most frequent words found in a reference corpus of Polish (IPI PAN Corpus) with one important exception made for public security terminology as well vocabulary currently used in this context (although not necessarily frequent in ordinary language). The initial PolNet was basically made of synsets built form simple (one word) nouns. The first stage of building PolNet ended with the resource amounting to some 11,700 synsets for over 20,300 word-senses (and 12,000 nouns). (The estimation of the effort invested in the development of the initial PolNet (for nouns) is 11 man-months of effective work.) Then this basic resource was extended with verb part which, in January 2012, was composed of over 1,500 synsets corresponding to some 2,900 word+meaning pairs for 900 most important Polish simple verbs. These works converge with noun-verb collocations development by Grażyna Vetulani and her team. Currently we proceed to the integration of verb-noun collocations into PolNet. This step is considered as a step torward the Lexicon Grammar of Polish (cf. the paper by Z. Vetulani and G. Vetulani (2012), as well as (Vetulani 2012)).

3.2.4.2. Top-level general ontology for POLINT-112-SMS (an option)

In the appendix to (Vetulani et al. 2010) by G. Taberski, the author presents initial works toward a top-ontology (top level ontology) which could cover the PolNet set of concepts (represented by synsets). Although PolNet used as ontology for POLINT-112-SMS appeared quite satisfactory, we decided to study the possibility to find (create) a general ontology (over POLINT) in order to make the potential integration of the system with other systems (for other languages) easier.[30]

Several existing proposals like SUMO [31], CYC [32], SOWA[33] and Top ontology for the EuroWordNet (cf. Vossen)[34] were considered with the conclusion that none of these is fully satisfactory. Ontologies SOWA and SUMO provide an useful partition of the universe into disjoint classes but are too general. CYC and EuroWordNet Top Ontology permit reasoning but leave many concepts without sharp classification. An initial proposal of a PolNet Top Ontology was inspired by all four mentioned above ontologies with however priority given to the EuroWordNet. It was assumed that the ontology will respect the natural asymmetry of the world (not all criteria are relevant to all concepts) and that the ontology should help reasoning. The general ontology should respect the conceptualization as it is reflected in Polish. The general separation line is between concepts related with various situations (such concepts are typically referred by verbs and predicative nouns) and all other kinds of concepts, typically denoted by common (non-predicative) nouns and adjectives.

The following criteria (called dimensions) help defining the classifying concepts. The dimensions are:

- compositionality (individuals vs. collections)
- function (the way in which the entity enters into relation with other entities)
- observability (real in time and space vs. abstract, existing only as an imagined entity)
- origin (natural vs. artifact)
- structure (the way entities are built)

[29] Verbs with four (or more) argument valency are, according to Grzegorczykowa (1998), rare in Polish. A systematic exceptions make *verbs of dislocation* (*Jan przestawia /move/ książki z/from/ szafy na/to/ półkę*).

[30] This project was initiated in the POLINT-112-SMS project, but is still in its initial phase. For the time being PolNet serves as the system ontology.

[31] Suggested Upper Merged Ontology (SUMO), http://www.ontologyportal.org/

[32] CYC Ontology, http://www.cyc.com/

[33] Cf. John F. Sowa (1999) in: (http://www.jfsowa.com/ontology/toplevel.htm)

[34] Piek Vossen in EuroWordNet General Document. http://www.illc.uva.nl/EuroWordNet/docs/GeneralDocDOC.zip; last access 10.04.2012

4. Concluding remarks

We claim to have shown, on a selected example, the role and diversity of language resources in the development of real-size systems with language competence in the area of public security. In the paper we presented several examples of language resource but we omitted several important factors, such as legal issues of acquisition, maintenance and visibility of sensitive data, especially in the perspective of commercialization of the results. These are hot aspects which merit separate studies.

5. Acknowledgements

The works presented in this paper have been partly supported by the Polish Ministry of Science and Higher Education, grant R00 028 02 (project „Polish Text Processing Technologies for the Public Security Oriented Applications" within the Polish Platform for Homeland Security). Many people contributed with their effort to the successful achievement of this project. I would like to thank the remaining co-authors of the post-project report (Vetulani et al., 2010)[35] (in Polish) from which much information of the current paper are reproduced.

References

Colmerauer, A. (1978): *Methmorphosis Grammars*, in: Bolc, L. (ed.): *Natural Language Communication with Computers*, Lecture Notes in Computer Science, 63, New York, N.Y., Springer.

Doroszewski, W. (ed.) (1958-1969): *Słownik języka polskiego,* Polska Akademia Nauk, PWN, Warszawa.

Dubisz, D. (ed.) (2006): *Uniwersalny słownik języka polskiego PWN*, (*Universal dictionary of Polish*, in Polish), 2nd edition, Warszawa: Wydawnictwo Naukowe PWN.

Fairon, C., Paumier, S. (2006): A translated corpus of 30,000 French SMS, in *Proceedings of LREC 2006,* Genova, ELRA/ELDA. Paris.

Filipkowski, W. (2004): *Zwalczanie przestępczości zorganizowanej w aspekcie finansowym*, Grupa Wolters Kluwer, Kraków.

Grzegorczykowa, R. (1998): *Wykłady z polskiej składni*, Warszawa, PWN.

Horàk, A., Pala, K., Obrebski, T., Rzepecki, P., Konieczka, Marciniak, Rambousek, A., Vetulani, Z. and Walkowska, J. (2007): DEB Platform tools for effective development of WordNets in application to PolNet, in: Vetulani, Z. (ed.) : *Proceedings of the 4th Language and Technology Conference, November 6-8, 2009*, Wyd. Poznańskie, Poznań, pp . 514-518.

Januszka, H., Gembara, S. (2005): *Elementy nowoczesnego zarządzania w policji*, Poznań.

Jędrzejko, E. (1998): *Słownik polskich zwrotów werbo-nminalnych, Zeszyt próbny*, Energeia, Warszawa.

Kubis, M. (2009): An access layer to a lexical database in POLINT-112-SMS, in: Vetulani, Z. (ed.*). Proceedings of the 4th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, November 6-8, 2009, Poznań, Poland.* Wyd. Poznańskie, Poznań , pp. 437-441.

Minsky, M. (1975), *A Framework For Representing Knowlege*. in: *The Psychology of Computer Vision*, McGraw-Hill, Nowy York, pp. 211-277,

Pereira, F. & Warren, D.H.D. (1980): Definite Clause Grammar for Language Analysis, in: *Artificial Intelligence*, vol. 13., 231 – 278 (also in Grosz, B.J., Jones, K.S., Webber, B.L. (eds.) (1986): *Readings in Natural Language Processing,* Morgan Kaufmann Publishers, Inc., pp. 101-124).

Polański, K. (red.) (1980-1992): *Słownik syntaktyczno-generatywny czasowników polskich*, t. I-IV, Ossolineum, Wrocław, 1980-1990, t. V, Instytut Języka Polskiego PAN, Kraków, 1992.

Rau, Zb. (2002): *Przestępczość zorganizowana w Polsce i jej zwalczanie* (rozprawa doktorska), Wyd. Zakamycze, Kraków.

Sowa, J. (1999): *Knowledge Representation: Logical, Philosophical, and Computational Foundations,* Brooks Cole Publishing Co., Pacific Grove, CA, ©2000.

Szymczak, M. (red.) (1981, 1995): *Słownik Języka Polskiego,* PWN, Warszawa

Vetulani, G.. (2000): *Rzeczowniki predykatywne języka polskiego. W kierunku syntaktycznego słownika rzeczowników predykatywnych*, Wydawnictwo UAM, Poznań.

Vetulani, G., Vetulani, Z., Obrębski, T. (2006): *Syntactic Lexicon of Polish Predicative Nouns*, N. Calzolari (ed.), Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, 24-26.05.2006, (Proceedings), ELRA, Paris, pp. 1734-1737.

Vetulani, Z. (1989): *Linguistic Problems in the Theory of Man-Machine Communication in Natural Language. A Study of Consultative Question Answering Dialogues. Empirical Approach*. Brockmeyer, Bochum.

Vetulani, Z. (1994): SWITCHes for making Prolog more Dynamic Programming Language", *Logic Programming, The Newsletter of the Association for Logic Programming,* vol 7/1, February 1994, p. 10.

Vetulani, Z. (1995a): *POLINT – system automatycznej interpretacji pytań w języku polskim i jego realizacja w PROLOGU,* in: Pogonowski, J. (ed.), *Euphonia i Logos. Księga pamiątkowa ofiarowana Profesor Marii Steffen-Batogowej i Profesorowi Tadeuszowi Batogowi,* Wydawnictwo Naukowe UAM, Poznań, 583–598.

Vetulani, Z. (1997): A system for Computer Understanding of Texts, in: R.Murawski, J. Pogonowski (eds), *Euphony and Logos* (Poznań Studies in the Philosophy of the Sciences and the Humanities, vol. 57) Rodopi, Amsterdam-Atlanta, 387-416.

Vetulani, Z. (2000): Electronic Language Resources for POLISH: POLEX, CEGLEX and GRAMLEX. In: M. Gavrilidou et al. (eds.), *Second International Conference on*

[35] The contributors to this report (Vetulani, Z., Marcinak, J., Obrębski, J., Vetulani, G., Dąbrowski, A., Kubis, M., Osiński, J., Walkowska, J., Kubacki, P., Witalewski, K. and G. Taberski) were active until the project end, but many other contributed as well during the project life (at least 42 people contributed at different stages, 30 of that number are members of academic staff - professors, PhD students and students, 12 are experts from the Polish Platform for Homeland Security).

Language Resources and Evaluation, Athens, Greece, 30.05.-2.06.2000, (Proceedings), ELRA, pp. 367-374.

Vetulani, Z. (2004): "Komunikacja człowieka z maszyną. Komputerowe modelowanie kompetencji językowej", Akademicka Oficyna Wydawnicza Exit: Warszawa.

Vetulani, Z. (2009): Natural Language Based Communication Between Human Users and Emergency Center in Critical Situations. A Short-Text-Message Based Decision Assistindg System POLINT-112-SMS, in: Vetulani, Z. (ed.) : *Proceedings of the 4th Language and Technology Conference, November 6-8, 2009*, Wyd. Poznańskie, Poznań.

Vetulani, Z., Kubis, M., Obrębski, T. (2010): *PolNet – Polish WordNet: Data and Tools*, in: Proceedings of LREC 2010. Valletta, Malta.

Vetulani, Z., Marciniak, J. (2000): *Corpus Based Methodology in the Study and Design of Systems with Emulated Linguistic Competence,* in: Dimitris N. Christodoulakis (red.), *Natural Language Processing – NLP 2000,* Lecture Notes in Artificial Intelligence, no 1835, Springer, 346-357.

Vetulani, Z., Marciniak, J., Konieczka, P., Walkowska, J. (2008): An SMS-based System Architecture (Logical Model) to Support Management of Information Exchange in Emergency Stuations. POLINT-112-SMS, In: Zhongshi Shi, E Mecier-Laurent, D. Lake (Eds.) *Intelligent Information Processing IV (Book Series: IFIP International Federation for Information Processing, Subject collection: Computer Science),* Volume 288/2008, Springer-Boston, pp. 240-253.

Vetulani, Z., Marcinak, J., Obrębski, T., Vetulani, G., Dabrowski, A., Kubis, M., Osiński, J., Walkowska, J., Kubacki, P., Witalewski, K. (2010): *Zasoby językowe i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego* (in Polish) (*Language resources and text processing technologies. POLINT-112-SMS as example of homeland security oriented application*), Adam Mickiewicz University Press: Poznań.

Vetulani, Z., Martinek, J., Obrębski, T., Vetulani, G. (1998b): *Dictionary Based Methods and Tools for Language Engineering*, Wyd. Naukowe UAM, Poznań.

Vetulani, Z., Obrębski, T. (2010): Resources for Extending the PolNet-Polish WordNet with a Verbal Component, in: Bhattacharyya, P, Fellbaum, Ch., Vossen, P. (eds.) *Principles, Construction and Application of Multilingual Wordnets. Proceedings of the 5th Global Wordnet Conference,*Narosa Publishing House: New Delhi, Chennai, Mumbai, Kolkata, pp. 325-330.

Vetulani, Z., Obrębski, T., Vetulani G. (2007): Towards a Lexicon-Grammar of Polish: Extraction of Verbo-Nominal Collocations from Corpora, in: *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference (FLAIRS-07),* AAAI Press (2007), Menlo Park, California, pp. 267-268.

Vetulani, Z., Obrębski, T., Vetulani, G. (2008): Verb-Noun Collocation SyntLex Dictionary – Corpus-Based Approach, Proceedings of LREC 2008, Marrakech, Morocco , ELRA, Paris.

Vetulani, Z., Walczak, B., Obrębski, T., Vetulani, G. (1998a): *Unambiguous coding of the inflection of Polish nouns and its application in the electronic dictionaries - format POLEX / Jednoznaczne kodowanie fleksji rzeczownika polskiego i jego zastosowanie w słownikach elektronicznych - format POLEX*, Adam Mickiewicz University Press, Poznań.

Vetulani, Z., Walkowska, J., Obrębski, T., Marciniak, J., Konieczka, P., Rzepecki, P. (2009): *An Algorithm for Building Lexical Semantic Network and Its Application to PolNet – Polish WordNet Project,* in: Z. Vetulani and H. Uszkoreit (Eds.): *Human Language Technology. Challenges of the Information Society,* LNAI 5603, Springer-Verlag Berlin-Heidelberg, pp. 369-381.

Vossen, P. (ed.) (2012), *EuroWordNet General Document. EuroWordNet* (LE2-4003, LE4-8328), Part A, Final Document, Version 3.

Walkowska, J. 2009., Gathering and Analysis of a Corpus of Polish SMS Dialogues, in: Kłopotek, M. A., Przepiórkowski, A., Wierzchoń, S. T., Trojanowski, K. (Eds.), *Challenging Problems of Science. Computer Science. Recent Advances in Intelligent Information Systems.* Academic Publishing House EXIT, Warsaw, pp. 145-157.

Walkowska, J. 2012., *Modelowanie kompetencji dialogowej człowieka na potrzeby jej emulacji w zarządzających wiedzą systemach informatycznych współpracujących z wieloma użytkownikami* (in Polish), PhD Thesis, UAM and IPI PAN, Poznań-Warsaw.