

A Framework for Identity Resolution and Merging for Multi-source Information Extraction

Milena Yankova^{*,†}, Horacio Saggion[†], Hamish Cunningham[†]

[†]Department of Computer Science
University of Sheffield
211 Portobello Street - Sheffield, England, UK
^{*}Ontotext Lab, Sirma Group
135 Tzarigradsko Chaussee - Sofia, Bulgaria
{milena,saggion,hamish}@dcs.shef.ac.uk

Abstract

In the context of ontology-based information extraction, identity resolution is the process of deciding whether an instance extracted from text refers to a known entity in the target domain (e.g. the ontology). We present an ontology-based framework for identity resolution which can be customised to different application domains and extraction tasks. Rules for identify resolution, which compute similarities between target and source entities based on class information and instance properties and values, can be defined for each class in the ontology. We present a case study of the application of the framework to the problem of multi-source job vacancy extraction.

1. Introduction

Ontology-based extraction (OBIE) is the process of identifying in text or other sources relevant concepts, properties, and relations expressed in an ontology. In the context of ontology-based information extraction, one fundamental problem to be addressed is that of identification and merging ontological instances extracted from multiple sources (the problem is known as ontology population in the Semantic Web community).

A consolidation procedure aims at identifying newly extracted (e.g. from text) facts and linking them to their previous mentions. Unlike classical information extraction (see (Grishman, 1997)) where the extracted facts are only classified as belonging to pre-defined types, identity resolution aims at establishing a reference link between an object residing in the system's knowledge base and its mention in context (e.g. text).

This paper presents a framework for Identity Resolution: the process of deciding if a particular fact extracted from text can be linked to identical/similar facts in the ontology. Recognising identical or similar information across different sources is of paramount importance and in particular can lead to improved extraction performance from single sources. Aggregation of extracted information allows for: (i) improving the completeness of the extraction; (ii) avoiding the extraction of incorrect information; (iii) adding a degree of trust to the extracted facts. Here we will introduce our Identify Resolution Framework (IdRF) which provides infrastructure for resolving identity of different classes of entities. The framework uses an ontology as an internal knowledge representation that provides detailed entity description formalism complemented with semantics. The framework is adaptable to different application domains and tasks.

The paper is organised as follows: Section 2 gives an overview of the framework. Its main components are briefly described with special emphasis on the first order probabilistic engine used for entity comparison. Sec-

tion 3 provides background information for our information extraction application use case and details implementation and evaluation results. The related work is discussed in Section 4 and conclusions and further work are given in Section 5.

2. Identity Resolution Framework

The framework is intended to provide a general solution to the identity problem that can be used within different applications regardless of their particular domain or type of entity which need to be resolved. The input is an entity together with its associated properties and values, the output is an integrated representation of the entity which will have new properties and values in the ontology.

A customisable identity criteria is in place to decide on the similarity between two instances. This criteria uses ontological operations and similarity computation between extracted and stored values which are weighted. The weighting criteria is specified according to the type of entity and the application domain.

2.1. Knowledge representation

The IdRF uses an ontology for internal and resulting knowledge representational formalism. The ontology not only contains the representation of the domain, but also known entities and properties. After identity resolution, the ontology Knowledge Base (KB) will contain entities with their full semantic description aggregated during the process.

As a side effect of this continuous updating of the KB, the identity criteria is refined, thus improving the identity resolution by both refining the evidence calculation and introducing new entities serving as identity goals. Details about the two effects are given below:

- The evidence calculation is refined when a new value, attribute, property or relation is added to an existing instance description. Then, the identity criteria for this instance is changed in order to reflect the newly available data adding new comparison restrictions. For ex-

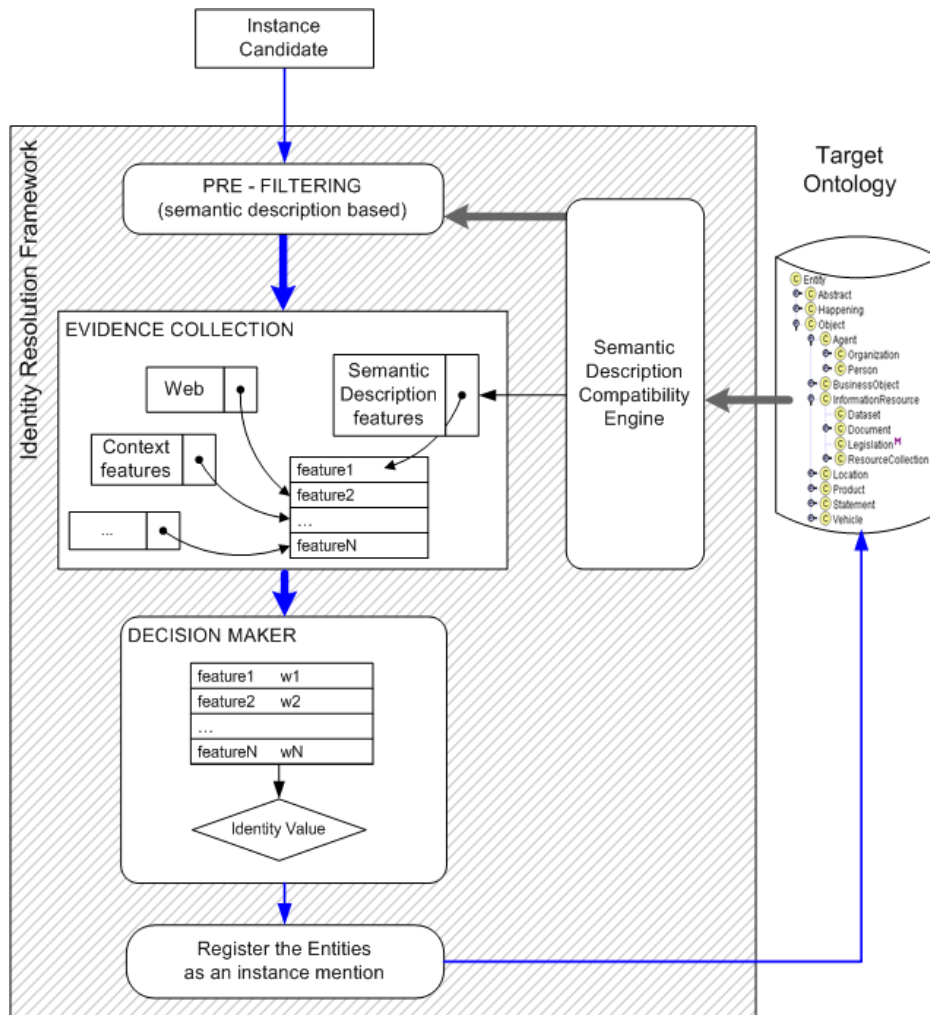


Figure 1: IdRF Architecture

ample if the person age is added to his/her description, the age restriction will be added a new identity criterion.

- New entities added to the knowledge base represent potentially new goals for resolution. They are created by insertion of entirely new objects to the KB. When entities are processed in a later stage, they have to be compared not only two the previously available entities but also to the newly added instances.

Current implementation of the IdRF is based on the PROTON (Terziev et al., 2005) ontology, which can be easily extended for any particular domain or specific task. The knowledge base that actually contains the ontology and the instances associated with it is stored in the semantic repository provided by KIM (Popov et al., 2004) that is based on OWLIM (Kiryakov et al., 2005) and Sesame.

2.2. IdRF Main Components

The IdRF framework receives an instance (e.g. type of instance and properties and values) and updates the ontology either asserting a new instance with its properties or updating an already existing instance. The IdRF architecture (see Figure 1) consists of four main stages.

- **Pre-filtering** - It filters out the irrelevant part of the ontology and forms a smaller set of instances similar to the source entity. It is intended to restrict the whole amount of ontology instances to a reasonable small number, to which the source entity will be compared. It can be regarded as pre-selection of ontology objects that are eligible to identification. The selected instances are potential target instances that might be identical to the source object; they already appear in the knowledge base and are somehow similar to the source object. Pre-filtering is realised by the Semantic Description Compatibility Engine (SDCE) which is described in details later on.
- **Evidence Collection** - It collects as much as possible evidence about the similarity between the source entity and each of the targets in the ontology. A set of similarity criteria is computed by SDCE by comparing corresponding attributes in the entity descriptions. Different comparison criteria are possible: some are based on string representation e.g. text edit distance, inverted frequency based matching; others can be web appearance, context similarity, etc.
- **Decision Maker** - Once all the evidences for different identity possibilities are collected, it concludes which

is the best identity match. It is this third stage that decides about the strength of the presented evidence and makes the decision. This module chooses the candidate favoured by the class model natively stored as part of the Class Model described in SDCE. The models are based on the weighting of evidences. The model can be easily tuned by domain experts.

- **Data Integration** - After the decision is made the incoming entity is registered to the ontology as a final stage in the IdRF. The source entity can be either new one or successfully identified with an existing instance. If the system is not able to find a reliable match, the incoming object is inserted as a new instance in the KB. In case it is associated with an existing instance, then the object description is added to the description of the identified KB instance. Thus, the result from the current identification is stored in the ontology and is used for further identity resolution of the next incoming objects.

2.3. Semantic Description Compatibility Engine

The main engine of this framework, Semantic Description Compatibility Engine (SDCE), is an implementation layer that provides access to the ontology. It is the backbone of the IdRF and it is used in both stages the Pre-filtering and the Evidence Collection of default identity criterion. SDCE creates class models that handle the specificity of different entity types presented as ontology classes. The instances of various classes differ in their meaning and type of their semantic descriptions, thus the class models describe different conditions for comparison during the identification process.

2.3.1. Class Models declaration

The engine is based on first order probabilistic logic calculus and each class model is expressed by a formula. Thus, each formula encodes the specificity of the corresponding class forming its model. All the formulas consist of predicates from a common pool of predicates, so several formulas may use the same predicate as part of their definitions. Each primitive predicate is implemented as Java class, so the set of predicates is extensible using Java programming language. It is essential that several formulas can use one and the same predicate as part of their definitions. This allows having a small set of reusable primitive predicates from which someone can compose complex formulas in a declarative way. However, the same predicate can be weighted differently according to its importance for the particular class modelled by different formulas. Formulas are composed from a set of primitive predicates combined with the usual logical connectives like like “AND”, “OR”, “NOT” and “IMPLICATION”. Different weights can be attached to each of the predicates in the formulas using the logical connective “&” and some real value from 0 to 1 (see Figure 2).

The parser of the SCDE associates formulas with specific classes in the ontology; it also supports rule inheritance between classes. So, the set of formulas can be easily expanded for a new class, when the ontology is extended

or the focus of the particular application of the IdRF is changed.

```
namespace:
rdf: "http://www.w3.org/1999/02/22-rdf-syntax-ns"
rdfs: "http://www.w3.org/2000/01/rdf-schema"
protons: "http://proton.semanticweb.org/2005/04/protons"
protonu: "http://proton.semanticweb.org/2005/04/protonu"
joci: "http://www.ontotext.com/2007/07/joci"

"protons:Entity":
  SameAlias()

"protonu:Company":
  let parentCond = Super() \& 0.7
      sectorCond =
        SameAttribute(<protonu:activeInSector>)
        aliasCond = SimilarCompanyAliases() \& 0.9
  in parentCond \& sectorCond \& aliasCond \&

"joci:Office":
  StrictSameAttribute("joci:hasURL") |
  OrganizationLD() |
  OrganizationCombine() \&
  StrictSameAttribute("joci:hasPostal") \&
  StrictSameAttribute("joci:hasSector")

"joci:Vacancy":
  let
    organisationCond =
      StrictSameAttribute("joci:hasMLID") |
      StrictSameAttribute("joci:hasContact")
  in SameAlias() \& organisationCond \&
  SameAttribute("joci:hasLocation") \&
  StrictSameAttribute("joci:hasRefNumber")
```

Figure 2: Example of rule definition

2.3.2. Class Models execution

There are two different ways of using the Class models by the SDCE depending on which component used the engine.

- **Pre-filtering** component finds those objects in the knowledge base that are possibly identical to the instance candidate and it uses SDCE to acquire them. The engine is able to compose a SeRQL query based on the input object and corresponding class model. Then it send the query to the semantic repository and returns the retrieved objects to the pre-filtering component.
- **Evidence Collection** component calculates the similarity between two objects based on their class model, which is expressed by a probabilistic logic formula. The result is a real value from 0 to 1, where value 0 means that the given entities are totally different and value 1 means that they are absolutely equivalent. Any value between 0 and 1 mean that these entities are equivalent but only with a specific confidence. Sometimes the similarity measure between two entities is based on the similarity between two other entities connected to the original one supported by usage of square bracket operator in the formula.

3. Case Study Evaluation

Here we present the evaluation of the identity resolution framework in the context of a job vacancy extraction task which is a multi-source extraction problem. It uses the Internet as a source – web-sites of companies, job-boards and

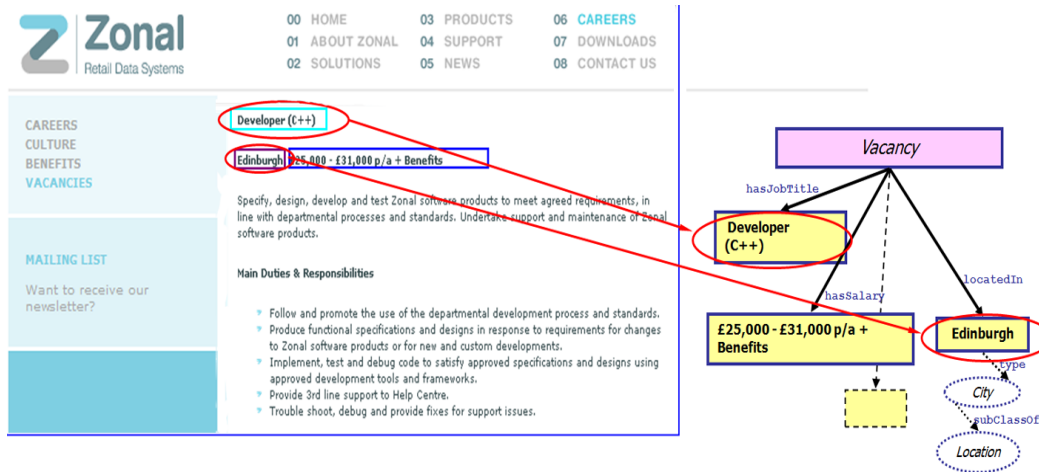


Figure 3: Single Vacancy extraction from a web page

recruitment agencies – where it is possible to rely on many sources to improve the extraction of the facts we search for. The proposed approach is to extract all available facts (in the job vacancy domain) from any single document, and then to combine/merge them on several levels to retrieve the most accurate facts, while at the same time filtering out wrong and redundant information.

3.1. Vacancy Extraction

The algorithm takes web-pages one by one and processes them separately, extracting listed vacancies. At this preliminary stage each page is pre-processed and certain types of named entities are recognised and annotated with respect to the ontology. After that, the set of extracted vacancies is further analysed to detect duplicates and finally inserted into the knowledge base.

Vacancy facts are defined by templates, which slots should be filled by concept instances in our KB. The extraction task consist on the extraction of the following information from text: JobTitle; ReportingTo; Job_Category; Job_Location; Location; Job_Reference; Job_Type; Salary; End_Date; Start_Date and Person. The proposed values for these attributes are named entities recognised by our system. Hence the extracted facts are actually a compilation of the attribute values in accordance with the domain constraints (see Figure 3). In Table 1, we present an evaluation of the extraction performance by slot. Overall, we have obtained F-score 87,4% (Precision 83,1% and Recall 92,3%) for single vacancy extraction.

3.2. Vacancy Merging

Once the vacancies are extracted the system proceeds with identification of those that are unique. For this purpose, we define Vacancy semi-equivalence is defined as follows: (i) equivalent “Vacancy Title” attribute values, or if one is a substring of the other, and (ii) the values of the rest of their attributes are semantically compatible according to the knowledge base, i.e. the two compared instances are connected with certain types of relation that is semantically consistent.

An example for such a relation is *subRegionOf* and we say that “locatedIn Wales” is comparable to “locatedIn UK”,

ATTRIBUTE	PRECISION	RECALL	F-MEASURE
JobTitle	0.87	0.86	0.85
ReportingTo	0.99	0.99	0.99
Job_Category	0.99	0.97	0.96
Job_Location	0.98	0.65	0.66
Location	0.89	0.93	0.88
Job_Reference	0.98	0.89	0.89
Job_Type	1	0.99	0.99
Salary	0.97	0.87	0.88
End_Date	0.99	0.93	0.93
Start_Date	0.98	0.98	0.89
Person	0.87	0.94	0.83

Table 1: Evaluation of single attributes extraction

since “Wales” is a *subRegionOf* of “UK”. What we achieve as a result of merging two vacancies is a new vacancy composed out of the most specific values among the two proposed values for each and for every attribute. All attribute values presented only in one of the merged facts are also taken. A very simple diagram on Figure 4 presents the choice of most specific values for “Vacancy Title” and “Vacancy Location” attributes presented as KB relations.

The motivation for the merging is the fact that one and the same vacancy is often promoted several times on a single (company) web-site. It starts from a list of vacant positions, followed by a very short description or a separate page with a detailed description of full vacancy details. The identity resolution is supported by the fact that all extracted position are offered in one and the same organisation. All this information gives us a chance to check the extracted facts and to collect all the available information provided by the employer when it is distributed on several pages.

Once having reliable single page IE results we investigate the redundancy phenomena. We took a sample of about 3k web sites and semi-automatically compared the extracted vacancies. Our experiment showed that about two thirds of the company web-sites have redundant job advertising. Moreover, the consolidation successfully reduces the number of facts to about 55% of the single page extracted results (see Table 2). The formal manual evaluation of the vacancy merging accuracy is given on Table 3.

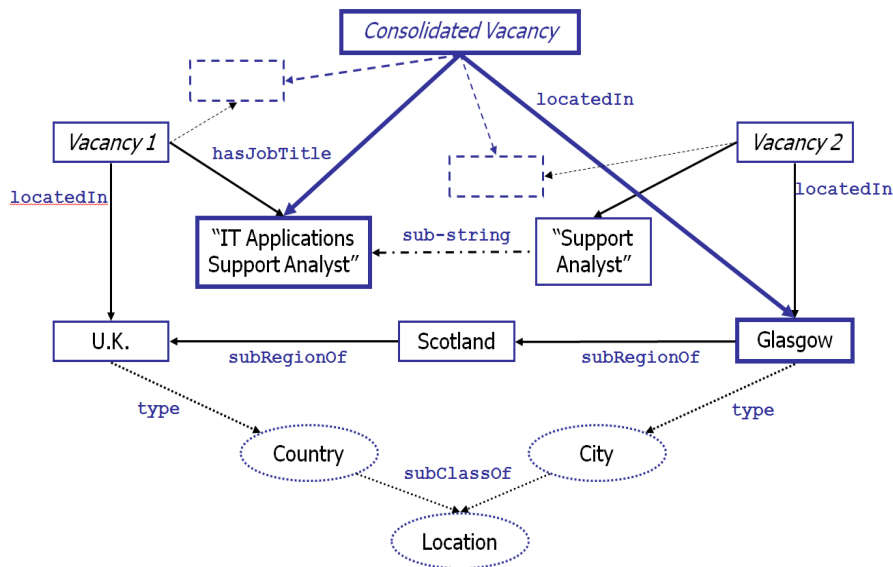


Figure 4: Example of consolidation of two Vacancy facts

STATISTICS	
web-sites with extracted <i>Vacancies</i>	2,922
web-sites with redundancy	2,171
<i>Vacancies</i> before merging	29,963
<i>Vacancies</i> after merging	16,592

Table 2: Redundancy Statistics

PRECISION	RECALL	F-MEASURE
0.82	0.89	0.85

Table 3: Evaluation of Intra-site vacancy merging

4. Related Work

Previous experiments in multi-source information extraction have been taken mainly in the area of Text Summarization, Databases and Co-reference Analysis. Bilenko and Mooney (Bilenko and Mooney, 2003) present a framework for duplicate detection using trainable measure of textual similarity (a learnable text distance function). Comprehensive survey about different methods used for de-duplication in database field is given by (Elmagarmid et al., 2007). However all the presented approaches are based on the string content of the corresponding field and hardly use even the fields' interdependence.

A notable aspect of using semantics for matching knowledge representation structures is presented by (Giunchiglia et al., 2004). The authors define Match as an operator that takes two graph-line structures and produces mappings among the nodes that correspond semantically to each other. However, the processing is based mainly on the node labels, even if their comparison is based on WordNet (Miller, 1994) and the graph structure is restricted to a tree. The IdRF proposed knowledge representation – ontologies – are already used for approaching the identity resolution problem. (Funk et al., 2007) present the advantages of semantically enhanced annotation for resolving co-references from different sources. Another example of using ontolo-

gies in this domain is the innovative work of (Klein et al., 2007) for extending standardised ontology description languages to unable approximation of instances. The authors introduce new “Rough Description Language” to represent and reason about similarity of instances.

From the natural language processing point of view, identity resolution has been addressed as a cross-document coreference task restricted to the problem of person coreference. Bagga and Baldwin (Bagga and Baldwin, 1998; Bagga and Biermann, 2000) used the vector space model together with summarization techniques to tackle the cross-document coreference problem. They use a Vector Space Model Disambiguation module and compute similarities between personal summaries (sentences extracted) for each pair of documents. Summaries having similarity above a certain threshold are considered to be about the same entity. Mann and Yarowsky (Mann and Yarowsky, 2003) use semantic information that is extracted from documents to inform a hierarchical agglomerative clustering algorithm. Semantic information here refers to factual information about a person such as the date of birth, professional career or education. Phan et al. (Phan et al., 2006) follow Mann and Yarowsky in their use of a kind of biographical information about a person. They use a machine learning algorithm to classify sentences according to particular information types. They compare information in automatically constructed person profiles by taking into account the type of the information. Entity identification is often addressed as author's name disambiguation in context of bibliographical records. In this context, Aswani et al. (Aswani et al., 2006) base their approach on web searches while looking for the author home pages, as well as papers' titles and abstracts. They mine information from the Web for authors including full name, personal page, and co-citation information to compute the similarity between two person names. Similarity is based on a formula which combines numeric features with appropriate weights experimentally obtained. Finally, Saggion (Saggion, 2008), studies the effect of dif-

ferent document contexts (e.g. full document, summary) and term representations (e.g. words, named entities) for entity clustering. An approach which uses named entities of type organisation to disambiguate person names proved to be very competitive.

5. Conclusions and Furure Work

We have presented a general framework for identity resolution which can be adapted to different ontology-based information extraction and ontology-population applications. We have also demonstrated and evaluated the application of the framework in the context of an ontology-based information extraction system. We are currently working on merging vacancies as well as organisations from sources different to corporate websites, e.g. job-boards. The approach taken uses consequential resolving of organisations followed by vacancy merging. Our future work will look into adapting the framework in the context of ontology population for business intelligence applications in financial risk management and internationalisation where target entities (e.g. companies, persons, locations) are extracted from multiple redundant sources requiring consolidation.

Acknowledgements

This work was partially supported by the EU-funded projects MUSING (IST-2004-027097) and MediaCampaing (027413).

6. References

- Niraj Aswani, Kalina Bontcheva, and Hamish Cunningham. 2006. Mining information for instance unification. In *5th International Semantic Web Conference (ISWC2006)*, Athens, Georgia.
- A. Bagga and B. Baldwin. 1998. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 79–85.
- A. Bagga and A. W. Biermann. 2000. A methodology for cross-document coreference. In *Proceedings of the Fifth Joint Conference on Information Sciences (JCIS 2000)*, pages 207–210.
- Mikhail Bilenko and Raymond J. Mooney. 2003. Employing trainable string similarity metrics for information integration. In *IJCAI-2003*, Mexico.
- Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vasilios S. Verykios. 2007. Duplicate record detection: A survey. Technical report, TKDE, January.
- Adam Funk, Diana Maynard, Horacio Saggion, and Kalina Bontcheva. 2007. Ontological integration of information extraction from multiple sources. In *International Workshop on Multi-source, Multi-lingual Information Extraction and Summarisation*.
- F. Giunchiglia, P. Shvaiko, and M. Yatskevich. 2004. S-match: an algorithm and an implementation of semantic matching. In *ESWS*, pages 61–75.
- R. Grishman. 1997. Information Extraction: Techniques and Challenges. In *Information Extraction: a Multidisciplinary Approach to an Emerging Information Technology*.
- Atanas Kiryakov, Damyan Ognyanov, and Dimitar Mano. 2005. Owlim a pragmatic semantic repository for owl. In *SSWS 2005, WISE*, USA.
- Michal C.A. Klein, Peter Mika, and Stefan Schlobach. 2007. Approximate instance unification using roughowl. In *Workshop on Uncertainty Reasoning for the Semantic Web (URSW)*.
- G. S. Mann and D. Yarowsky. 2003. Unsupervised personal name disambiguation. In W. Daelemans and M. Osborne, editors, *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003)*, pages 33–40. Edmonton, Canada, May.
- George A. Miller. 1994. Wordnet: a lexical database for english. In *HLT '94*, USA.
- X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. 2006. Personal name resolution crossover documents by a semantics-based approach. *IEICE Trans. Inf. & Syst.*, Feb 2006.
- Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, and Angel Kirilov. 2004. Kim - a semantic platform for information extraction and retrieval. In *Journal of Natural Language Engineering*. Cambridge University Press.
- H. Saggion. 2008. Experiments on semantic-based clustering for cross-document coreference. In *International Joint Conference on Natural Language Processing*, Hyderabad, India, January. AFNLP.
- Ivan Terziev, Atanas Kiryakov, and Dimitar Mano. 2005. Base upper-level ontology (bulo) guidance. Technical Report Deliverable 1.8.1, SEKT project, UK, July.