Proceedings of the
ELRA Workshop on Evaluation
Looking into the Future of Evaluation:
When automatic metrics meet task-based and performance-based approaches

Edited by Victoria Arranz, Khalid Choukri, Bente Maegaard and Gregor Thurmair

Marrakech, Morocco

27 May 2008

Workshop Programme

09:00 Welcome and Introduction

09:15 ***Technology Advancement has Required Evaluations to Change Data and Tasks -- and now Metrics***
Mark Przybocki; NIST, USA

10:00 ***Explicit and Implicit Requirements of Technology Evaluations: Implications for Test Data Creation***
Lauren Friedman, Stephanie Strassel, Meghan Lammie Glenn; Linguistic Data Consortium, USA

10:35 Coffee break

11:00 ***Automated MT Evaluation for Error Analysis: Automatic Discovery of Potential Translation Errors for Multiword Expressions***
Bogdan Babych, Anthony Hartley; Centre for Translation Studies, University of Leeds, United Kingdom

11:35 *Discussion*

12:15 ***Reference-based vs. Task-based Evaluation of Human Language Technology***
Andrei Popescu-Belis; IDIAP Research Institute, Switzerland

12:50 ***FEIRI: Extending ISLE's FEMTI for the Evaluation of a Specialized Application in Information Retrieval***
Keith J. Miller; The MITRE Corporation, USA

13:25 Lunch

14:30 *Discussion*

15:15 ***Evaluating a Natural Language Processing Approach in Arabic Information Retrieval***
Nasredine Semmar, Laib Meriama, Christian Fluhr; CEA, LIST, Laboratoire d'ingénierie de la Connaissance Multimédia Multilingue, France, NewPhenix, France

15:50 Coffee break

16:20 ***A Review of the Benefits and Issues of Speaker Verification Evaluation Campaigns***
Asmaa El Hannani, Jean Hennebert; Department of Computer Science, University of Sheffield, UK, HES-SO, Business Information Systems, Switzerland and University of Fribourg, Switzerland

16:55 ***Field Testing of an Interactive Question-Answering Character***
Ron Artstein, Sudeep Gandhe, Anton Leuski and David Traum; Institute for Creative Technologies, University of Southern California, USA

17:30 Discussion and conclusions

18:15 Close

Workshop Chairing Team

Gregor Thurmair, *Linguatec Sprachtechnologien GmbH, Germany* - **chair**
Khalid Choukri, *ELDA - Evaluations and Language resources Distribution Agency, France* – **co-chair**
Bente Maegaard, *CST, University of Copenhagen, Denmark* – **co-chair**

Organising Committee

Victoria Arranz, *ELDA - Evaluations and Language resources Distribution Agency, France*
Khalid Choukri, *ELDA - Evaluations and Language resources Distribution Agency, France*
Christopher Cieri, *LDC - Linguistic Data Consortium, USA*
Eduard Hovy, *Information Sciences Institute of the University of Southern California, USA*
Bente Maegaard, *CST, University of Copenhagen, Denmark*
Keith J. Miller, *The MITRE Corporation, USA*
Satoshi Nakamura, *National Institute of Information and Communications Technology, Japan*
Andrei Popescu-Belis, *IDIAP Research Institute, Switzerland*
Gregor Thurmair, *Linguatec Sprachtechnologien GmbH, Germany*

Contents

Introduction

Automatic methods to evaluate system performance play an important role in the development of a language technology system. They speed up research and development by allowing fast feedback, and the idea is also to make results comparable while aiming to match human evaluation in terms of output evaluation. However, after several years of study and exploitation of such metrics we still face problems like the following ones:

- they only evaluate part of what should be evaluated
- they produce measurements that are hard to understand/explain, and/or hard to relate to the concept of quality
- they fail to match human evaluation
- they require resources that are expensive to create

etc. Therefore, an effort to integrate knowledge from a multitude of evaluation activities and methodologies should help us solve some of these immediate problems and avoid creating new metrics that reproduce such problems.

Looking at MT as a sample case, problems to be immediately pointed out are twofold: reference translations and distance measurement. The former are difficult and expensive to produce, they do not cover the usually wide spectrum of translation possibilities and what is even more discouraging, worse results are obtained when reference translations are of higher quality (more spontaneous and natural, and thus, sometimes more lexically and syntactically distant from the source text). Regarding the latter, the measurement of the distance between the source text and the output text is carried out by means of automatic metrics that do not match human intuition as well as claimed. Furthermore, different metrics perform differently, which has already led researchers to study metric/approach combinations which integrate automatic methods into a deeper linguistically oriented evaluation. Hopefully, this should help soften the unfair treatment received by some rule-based systems, clearly punished by certain system-approach sensitive metrics.

On the other hand, there is the key issue of « what needs to be measured », so as to draw the conclusion that « something is of good quality », or probably rather « something is useful for a particular purpose ». In this regard, works like those done within the FEMTI framework have shown that aspects such as usability, reliability, efficiency, portability, etc. should also be considered. However, the measuring of such quality characteristics cannot always be automated, and there may be many other aspects that could be usefully measured.

This workshop follows the evolution of a series of workshops where methodological problems, not only for MT but for evaluation in general, have been approached. Along the lines of these discussions and aiming to go one step further, the current workshop, while taking into account the advantages of automatic methods and the shortcomings of current methods, should focus on task-based and performance-based approaches for evaluation of natural language applications, with key questions such as:

- How can it be determined how **useful** a given system is for a given task?
- How can focusing on such issues and combining these approaches with our already acquired experience on automatic evaluation help us develop new metrics and methodologies which do not feature the shortcomings of current automatic metrics?
- Should we work on hybrid methodologies of automatic and human evaluation for certain technologies and not for others?
- Can we already envisage the integration of these approaches?
- Can we already plan for some immediate collaborations/experiments?

What would it mean for the FEMTI framework to be extended to other HLT applications, such as summarization, IE, or QA? Which new aspects would it need to cover?

Workshop Programme and Audience Addressed

This full-day workshop is intended for researchers and developers on different evaluation technologies, with experience on the various issues concerned in the call, and interested in defining a methodology to move forward.

The workshop features one invited talk, submitted papers, and will have ample time for discussion on future developments and collaboration.

Explicit and Implicit Requirements of Technology Evaluations: Implications for Test Data Creation

Lauren Friedman, Stephanie Strassel, Meghan Lammie Glenn
Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia, PA 19103
{lf, strassel, mlglenn}@ldc.upenn.edu

Abstract

A multitude of approaches, methodologies and metrics exist for evaluating the performance of technologies like machine translation, speech recognition and information extraction. While metrics vary widely in their assumptions about what is being tested and how it should be measured, most technology evaluations rely crucially on a carefully constructed test data set that is both accurate and fully expressive of the phenomena being evaluated. Within this context, this paper explores some of the challenges of creating reference data for technology evaluations, highlighting many of the decisions and judgments that must be made with regard to data selection, difficulty, annotation, and quality. We discuss not only the fully articulated expectations for test data, but also the hidden assumptions and implicit requirements that affect test set creation. We use the GALE Machine Translation task as a case study in discussing these issues, occasionally drawing examples from other evaluations to illustrate various aspects of the problem.

1. Introduction

A multitude of approaches, methodologies and metrics exist for evaluating the performance of technologies like machine translation, speech recognition and information extraction. While metrics vary widely in their assumptions about what is being tested and how it should be measured, most technology evaluations rely crucially on a carefully constructed test data set. While some metrics require post-hoc manual assessment of system performance, even automatic metrics like BLEU and METEOR assume the existence of one or more gold standard references against which system performance can be compared. Different metrics vary in their requirements about the completeness of the reference data or the extent to which multiple "right answers" can exist, but nearly all assume that the reference data is both accurate and fully expressive of the phenomena being evaluated.

Within this context, this paper explores some of the challenges of creating reference data for technology evaluations. We use the GALE Machine Translation task as a case study in discussing these issues, occasionally drawing examples from other evaluations to illustrate various aspects of the problem.

On the surface, creation of test data for a task like machine translation is straightforward: take the set of evaluation documents and manually translate them. But like any task involving human judgment, "translation" is not a monolithic task and there are multiple decision points along the way. In the sections that follow, we discuss several of these decision points, considering not only the fully articulated requirements for test data – the type stated in an evaluation plan – but also hidden assumptions and implicit requirements that are equally important in constructing appropriate data for evaluation.

2. Data Selection

First, we consider the question: what data is appropriate for inclusion in the test set? From the perspective of a system developer, a good test set is one whose profile is reasonably similar to that of available training and devtest data. Project sponsors and customers, on the other hand, may expect systems to handle previously unseen challenges.

The ability of data creators to balance these two opposing requests is limited by the pre-determined collection epoch for each evaluation. Irrespective of stakeholders' expectations, the profile of the final test set will be dictated at least partially by the pool of available data.. Some features of the evaluation set – its topic coverage, for example – will be necessarily distinct from what is found in training and devtest data. Thus the specification of a test epoch can automatically add novel challenges to the evaluation. Challenges introduced by the epoch constraint are features of the available data pool and outside of the control of data creators. While a narrowly defined evaluation epoch can increase difficulty, it also limits the range, scope, and variability possible within a test set.

Data creators are often in the difficult position of balancing these conflicting requirements and limitations when selecting data for inclusion in the test set. To make things still more challenging, the "profile" of any given set of data is highly multidimensional, including such components as language, dialect, genre, source, structure, topic, time epoch, document length, segment length, lexical variation, difficulty, etc. While some of these components (summarized in Table 1) are clear cut and unambiguous (e.g. document length), others are less well-defined.

	Unambiguously Specified in Typical Eval Plan?	Directly Measurable / Testable During Eval Set Creation?
Language	m	y
Dialect	n	m
Genre	m	m
Source	y	m
Topic	n	m
Epoch	y	y
Document Structure	n	m
Source Data Format	m	y
Encoding	y	y
Doc Length	y	y
Segment Length	y	y
Lexical Variety	n	m
Linguistic/Structural Complexity (e.g. syntax)	n	n
Overall Difficulty	n	n

Table 1: A subset of data features.

For the NIST Open Machine Translation Evaluation, for example, the evaluation plan developed by NIST included clear direction on goals, training conditions, test data, file formats, and performance metrics (NIST, 2008). Such a detailed evaluation plan is valuable not only for participating sites, but for data creators as well.

The sheer number of data variables and types, however, makes it impossible to fully account for the effect of individual components and the various interactions among them. Data creators endeavor to build a test set according to specifications described in an evaluation plan. But the "ideal" balance of components remains elusive since the impact of certain factors is not yet known – and in some cases cannot be fully known – and the various components are often non-orthogonal.

While all efforts are made to meet any explicit expectations, blindly following only the expectations specified in an evaluation plan does a disservice to the program. Without understanding finer points about the data itself, the goals of the evaluation, and the design of the evaluation metrics, data creators might make choices during test set construction that have unintended consequences. Having detailed expectations stated explicitly in an evaluation plan is essential, but it's not enough. Since decisions on subtler points of the data will always be necessary, data creators must have well-rounded knowledge of all aspects of an evaluation.

For instance, in a typical translation task we assume that the source and target languages are constant between the training and test data partitions. Confirming the language of a given set of documents seems trivial, but there can be hidden challenges. For example, in the case of Arabic, some informal genres like weblogs may show a substantial amount of colloquial Arabic mixed with Modern Standard Arabic. The amount of dialect mixture and the particular dialects represented can vary widely from one source to the next, from one individual document to the next, and even within a single document. A test set unwittingly selected from dialect-heavy documents, sources or genres may be significantly more challenging than the training data.

3. Test Set Difficulty

The question of test set difficulty is particularly important for evaluations that include "go/no-go" performance targets, such as the DARPA GALE program, since the program's continuation depends in part on the ability of translation systems to meet these pre-defined targets. Fair and accurate quantification of performance and measurement of progress require a test set whose make-up is carefully controlled and fully intentional. In GALE, unsurprisingly, considerable effort is devoted to selecting an annual test set whose difficulty is closely matched to the previous year's test set. The selection process begins with human annotators reviewing a pool of candidate documents, making judgments about language, dialect, genre and topic category; annotators also give a preliminary document difficulty rating on the Interagency Language Roundtable (ILR) scale (Clifford et al, 2004). The selection may be further refined by a series of automatic diagnostics to calculate log-perplexity and tri-gram hit-rate for documents in the candidate pool, in order to identify those that are outliers when compared to the rest of the selection pool and/or previous MT evaluation sets. TER (translation edit rate [Przybocki, Sanders, & Le, 2006]) may also be calculated for translated candidate documents as another measure of test set difficulty.

This approach – with several stages of data analysis and filtering – ensures that as many components as possible are known factors when building the final test set. However, even with all data features available to aide the selection process, the measure of "difficulty" is by no means straightforward. MT systems have different weak points and will find different areas of the data especially challenging. Assessing disparate data components when constructing a test set is important in order to balance test difficulty for all evaluation participants, but also to provide evaluation coordinators and sponsors with a reliable metric for gauging actual performance improvements over time.

The continued growth of multi-year programs, such as GALE, is somewhat constrained by the need for consistent test data; since the performance targets are set from the beginning, the difficulty of test sets for all phases must match that of the first in order to reliably measure progress. For example, if Phase 1 data is found to be too difficult, that inflated level of difficulty will be preserved for the duration of the program; otherwise, any conclusions drawn from trends in performance over subsequent phases will be untenable. Although the data selection process for GALE has become lengthier and more complex with each year's evaluation, there will always be unknowns, and matching difficulty from one phase to the next remains a significant challenge.

A "progress set" offers one alternative approach to the problem of measuring improvement against a test set that is different each year. While GALE does not include a

progress set, the DARPA EARS Program introduced the idea of designating a subset of evaluation data that remains blind for the duration of a program (Strassel, 2004). While this progress set introduces a new list of challenges – including the long-term sequestration of data – it does offer a fixed yardstick for the measure of progress over time. Whether the potential benefits of a progress set outweigh its added costs and complications is an open question.

4. Data Annotation and Quality

Assuming the question of test data selection has been settled, the selected data is typically annotated in some fashion – transcribed, translated, tagged for entities – to create the gold standard reference. Here too there are a multitude of challenges for the data creator in ensuring the test set is well-matched to the evaluation. The goals of the evaluation must be utterly explicit in terms of what is being measured and how; it is also important for data creators to understand the desired application for the technology being evaluated. All of this has a bearing on what the reference should consist of and how it should be created, but often these goals are only defined in the broadest of terms.

In a translation task for instance, the goal is known to be the production of "high quality" MT. But how important is fluency versus completeness or precision of meaning? Ideally, all of these features are present in a high-quality translation, but – in reality – they are often at odds. All of the possible MT goals that the FEMTI framework (King, Popescu-Belis, & Hovy, 2003) identifies require different emphases during the creation of the evaluation set. The desired use of the MT technology and the context within which it will be applied shape the priorities of the system developers and the evaluators, and these same details must also guide the data creators.

If the goal of the evaluation is to generate readable translations, the data creator might be tempted to heavily emphasize fluency when producing the reference translations. But a measure such as readability is difficult to quantify and almost entirely dependent upon the intended use of the data. A domain expert might prefer that subtleties of meaning be preserved even at the expense of fluency, while a novice reader might reverse these preferences.

The target consumer should guide these choices but is often an unknown quantity. Even when the audience is known, its needs are not always fully articulated or understood. And if the consumer of the translations is not a human at all but another downstream application (information retrieval, summarization, entity extraction, etc.), readability becomes something else entirely; both fluency and semantic accuracy could become secondary concerns if the preservation of word order, for example, is a requirement for a downstream task. Even when the technology goal itself is straightforward, a brittle evaluation paradigm with too many competing requirements will make the data creation task unmanageable.

The question of quality is central in test data creation. The term gold standard implies that the resulting resource is the best that humans can produce. But while there are several ways translation quality can be measured, there is always a subjective component. A universally accepted objective standard for human translation quality is probably untenable since, at least in the context of technology evaluations, translation quality must always be judged in terms of its intended use. The translation that will be most useful to the target consumer and the translation that will evaluate an MT system most fairly are not one and the same. Consequently, as with the question of data selection, the opinions of system developers and project sponsors do not always dovetail on what constitutes high quality data.

In addition to the lack of consensus in defining data quality, hidden assumptions can impede appropriate creation of a gold standard for any given evaluation. What kinds of humans, with what skills or training and with what kind of infrastructure, are expected to produce the gold standard? For instance, a run of the mill commercial translation will represent the work of one, or perhaps two, translators. But for many evaluation paradigms, the gold standard translation represents the collective effort of a much larger team; in the GALE program, gold standard translations require a series of manual passes by at least six individuals:

> 1) source-language dominant bilingual translator produces a preliminary translation emphasizing accuracy;
> 2) target-language dominant bilingual translator revises the translation to improve fluency;
> 3) source-language dominant bilingual annotator checks translation for errors and omissions;
> 4) source-language dominant bilingual senior annotator checks for remaining errors, improves fluency, corrects and standardizes named entities;
> 5) target-language dominant bilingual annotator improves fluency and adds translation variants where required;
> 6) target-language monolingual annotator reviews for fluency and flags questionable regions.

By any reasonable definition, the GALE gold standard translations can be said to be high quality, but the quality is in some ways artificial. The final references, as the product of a carefully constructed team, are far beyond the scope of what a single human translator could generate. Thus the MT is not scored against a human translation that could in any way be considered representative, but against a composite translation that is polished an almost unreasonable number of times.

This laborious process for gold standard creation was defined with the specific requirements of the GALE evaluation firmly in mind. The GALE evaluation metric is HTER, defined as the minimum number of edits one must make to the MT output so that it has the same meaning as the gold standard reference and is equally understandable (Przybocki, Sanders & Le, 2006). Given this metric, the gold standard references for GALE have properties that are not required for many other MT evaluations, and are not frequently found in run of the mill commercial translations. For example, when the source

text's meaning is ambiguous (e.g., verb tense is not expressed in Chinese), variants are added to the gold standard translation. In a standard translation, a translator would resolve ambiguities based on context and judgment, but the GALE gold standards require that the presence of this ambiguity is carefully preserved. Similarly, idioms are translated both literally and figuratively. The final references are meant to be not only fluent and accurate, but also completely inclusive of all reasonable interpretations of the source. This approach seeks to address the "multiple correct answers" problem of translation and ensure the fairest possible evaluation of MT systems.

Another dimension of test set quality is the consistency of the reference annotation. For annotations that require multiple passes by multiple judges like the GALE gold standard translations described above, it is difficult to imagine what "consistency" would mean, or how it could be measured. With a metric like edit distance, a high level of consistency is not really possible, expected, or even desirable. The multiple passes on GALE evaluation translations, for example, actually take inconsistency as a baseline assumption; each stage of quality control is intended to produce output that differs from – and improves upon – the previous stage. The expectation of this approach is not consistency between annotators, but rather the consistency of this group as a whole. While the group may not be internally consistent, the consistency between this group and other similarly-constructed groups can be expected to be greater than the consistency between two individuals.

Other tasks are superficially more straightforward, like orthographic transcription of audio data. As part of the DARPA EARS program in 2004, LDC undertook a careful study of inter-transcriber consistency, using the RT-03 English current test set (Strassel, 2004). Each evaluation file was transcribed by two annotators working independently, and the resulting transcripts were compared using the standard scoring software developed by NIST for the program's speech-to-text evaluation (NIST, 2004). While consistency was good, it was by no means perfect: the broadcast news genre showed a word disagreement rate of 1.1%, while conversational telephone speech showed 4.3% disagreement. These numbers are quite low in absolute terms, but given go/no-go performance targets of 5-10% word error rate and better for STT systems, it is critical to establish a baseline for human "performance". For more complex tasks, consistency rates are typically lower.

Performance targets are being set higher and higher; can machine error rate be reasonably expected to drop as low as – or lower than – the rates of human variation? Or should systems only be expected to perform somewhere within the range of typical human error? The urgency of resolving this issue rises as the gap between machine performance and human consistency narrows with each evaluation campaign.

5. Conclusion

The challenges for test set creation discussed in the sections above are not unique to evaluation data; they are relevant to any linguistic resource created for a particular purpose. With evaluation data, however, the stakes are typically higher and so the pressure on data creators is more intense. This is often coupled with a shorter timeline for developing evaluation data (compared to training data), which can be quite challenging given the primary emphasis on quality and the increased importance of consistency. As a result, the overall cost for test data creation is typically many times higher than training data created for the same evaluation. For GALE MT for instance, gold standard references are roughly ten times more costly (in dollars and time) than training data references, even though the training data can also be characterized as high quality.

The process for creating training data, though the end product is certainly high quality, only minimally resembles the gold standard creation process – even within the same program. While test set creation is so intensive precisely because the stakes are so high and the margin for error is so low, the effect of the schism between these two approaches to data creation needs to be further interrogated. The protocols for evaluation data could not reasonably be applied to training data, given the high volumes required of the latter. But what is the significance, if any, of training systems on data that is constructed differently, with different quality standards, than the test data that will ultimately be used to evaluate them?

The creation of gold standard references is so resource-intensive that even scaling up or supporting multiple evaluations at once becomes an inordinate challenge. The significantly higher costs of test set creation are only justifiable if higher quality can be shown to correlate with fairer evaluation – a correlation that is nearly impossible to prove. The high cost of evaluation data creation further underscores the importance of clearly defining the goals of the evaluation, fully informing data creators of program requirements, and then closely matching the test data to these needs and goals.

6. Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

7. References

Clifford, Ray, Neil Granoien, Douglas Jones, Wade Shen, & Clifford Weinstein (2004). "The effect of text difficulty on machine translation performance: a pilot study with ILR-rated texts in Spanish, Farsi, Arabic, Russian and Korean." In: *LREC-2004: Fourth International Conference on Language Resources and Evaluation*, Proceedings, Lisbon, Portugal, 26-28 May 2004; pp.343-346.

King, M., Popescu-Belis, A. and Hovy, E. 2003. "FEMTI: creating and using a framework for MT evaluation." In:

AMTA (2003), 224-231.

National Institute for Standards and Technology (2004). NIST RT-04 Spoken Language Technology Evaluation. http://www.nist.gov/speech/tests/rt/rt2004/fall/index.htm

National Institute for Standards and Technology (2008). *The 2008 NIST Open Machine Translation Evaluation Plan 2.4.* http://www.nist.gov/speech/tests/mt/2008/doc

Przybocki, Mark, Gregor Sanders, & Audrey Le (2006). "Edit distance: a metric for machine translation evaluation." In: *LREC-2006: Fifth International Conference on Language Resources and Evaluation*. Proceedings, Genoa, Italy, 22-28 May 2006; pp.2038-2043

Strassel, Stephanie (2004). "Linguistic Resources for Effective, Affordable, Reusable Speech-to-Text." In: *LREC-2004: Fourth International Conference on Language Resources and Evaluation*. Proceedings, Lisbon, Portugal, 26-28 May 2004.

Automated MT evaluation for error analysis: automatic discovery of potential translation errors for multiword expressions

Bogdan Babych, Anthony Hartley

Centre for Translation Studies, University of Leeds

Leeds, LS2 9JT, UK

b.babych@leeds.ac.uk, a.hartley@leeds.ac.uk

Abstract

We describe an on-going research project aimed at automatic detection of MT errors using state-of-the-art MT evaluation metrics, such as BLEU. Currently, these automated metrics give only a general indication of translation quality at the corpus level, and cannot be used directly for identifying gaps in coverage of MT systems. Our methodology uses automatic detection of frequent multiword expressions (MWEs) in sentence-aligned parallel corpora and computes an automated evaluation scores for concordances generated for such MWEs which indicates whether a particular expression is systematically mistranslated in the corpus. The method can be applied both to source and target MWEs, indicating whether MT can successfully deal with source expressions, or whether certain frequent target expressions can be successfully generated. The results can be useful for systematically checking the coverage of MT systems in order to speed up the development cycle of rule-based MT. This approach can also enhance current techniques of finding translation equivalents by distributional similarity and of automatically deriving specifications and rewrite rules for MT-tractable language.

1. Introduction

Automated MT evaluation methods – such as BLEU, NIST and Meteor – have been shown to be useful for monitoring progress in MT development, for parameter optimisation of statistical systems and, in some controlled circumstances, for comparing the performance of different MT systems. All such MT evaluation experiments rely on a corpus of human translations which are used as a reference for the MT output. Automated evaluation scores correlate with human scores and correctly establish ranking of systems only if this corpus is relatively large, i.e., more than 6,000-7,000 words (Estrella et al. 2007; Babych et al. 2007b). Smaller samples of data are too noisy for reliably predicting a system's performance, since individual lexical mismatches between MT output and human reference are not informative on their own: they can be attributed either to translation errors or to choices of different legitimate translation variants. While human judgements are meaningful at any granularity for which they are generated (the levels of syntactic constituent, sentence, paragraph, text and corpus as a whole), automated scores are generally not meaningful at any level below corpus. As a result, automated evaluation scores are currently uninformative for error analysis tasks – specifically, for discovering typical translation errors and prioritising them for the purposes of MT development – since they give only very a general, 'birds-eye' view of MT performance.

Moreover, MT developers are often less interested in such non-specific performance figures than in more detailed analysis and ranking of typical problems for their MT system whose resolution will improve the system's performance generally. As a result, developers of industry-standard (especially rule-based) systems consider these core automated evaluation metrics to be of little help in the MT development cycle (Thurmair 2007), noting that they are not designed to provide direction to R&D (Miller and Vanni 2005). Although human evaluation scores can be much more useful in this respect, they are expensive to obtain and not available for significantly large corpora. Thus it is difficult to rely on them for determining the range, frequency and seriousness of errors and, especially, for monitoring the progress of an MT system over time.

From this perspective, the challenge for automatic MT evaluation research is to develop a methodology which is suitable for differentiated and fine-grained error analysis along the lexical, grammatical and stylistic dimensions. Our paper reports on an on-going project for automatically discovering and ranking errors in translating multiword expressions (MWEs). While at this stage our methodology targets only the lexical dimension, this is proposed as a useful step towards more practical MT evaluation for developers and users of state-of-the-art MT systems.

2. Methodology

Our method is based on automatic evaluation of the translation of concordances for frequent MWEs extracted from aligned corpora. The methodology includes the following stages.

1. We generate automatically frequency-ranked lists of continuous and discontinuous MWEs, using the approach described in (Babych et al. 2007a), which relies on a combination of part-of-speech and frequency filters. We modified this approach in order not to depend on morphological annotation, therefore making it knowledge-light and language-independent. The idea came from an observation that part-of-speech filters typically prevent the appearance of function words on one or both edges of MWEs. For example, *visual processing *to / *in / *and* are filtered out, leaving only *visual processing* as a candidate MWE, which is selected if it passes a

$$logIDF = \log\left(\frac{N}{df_i}\right) > 1$$

certain frequency threshold in a corpus. Instead of using such a part-of-speech filter, we filter by log IDF scores:

where N is the number of texts in corpus and df_i is the number of texts where the word$_i$ is found. The threshold logIDF > 1 yields a relatively good distinction between content and function words. Function words are included within candidate continuous MWEs (a productive pattern, especially in Romance languages) but are excluded from discontinuous MWEs.

2. We generate automatically concordances for the most frequent MWEs in a sentence-aligned MT evaluation corpus, such as DARPA-94 (White et al. 1994). The concordances contain the MWEs themselves and several words in their local context. Thus the concordances can be viewed as sub-corpora selected by a specific MWE, intended to characterise the successfulness of their translation by MT. There are two possible scenarios for generating such concordances.
   a. If an operational MT system is available to the evaluators, then concordances can be generated on the Source Language (SL) side and submitted to the system for translation. These outputs will show whether the MT system can successfully deal with specific SL MWEs.
   b. If no MT system is available to the evaluators (only the texts are available, as is the case with the DARPA-94 corpus), then the concordances can be generated on the Target Language (TL) side from the human reference translations. These concordances show another aspect of MT performance: whether a system can successfully generate the most frequent TL MWEs.

   Scenario 2(a) corresponds to the most common commercial evaluation situation, whereas scenario 2(b) may occur in the context of meta-evaluation.
3. We compute a family of standard automated evaluation scores, including BLEU, for each of the concordances. In scenario 2(a) the test and the reference are determined in the usual way: the MT-translated concordances become a test set, and their corresponding aligned sentences from the human translations become the reference. Since we do not use word alignment (which may be too noisy), the whole segments aligned with the concordance segments become the reference. So reference texts may now be much longer than tested concordances. This, however, is not a problem for BLEU, which is an asymmetric, Precision-based metric: with the brevity penalty switched off, BLEU is only interested whether a test file contains any spurious items which are not found in the reference. Therefore, the reference text can be arbitrarily large. In scenario 2(b) concordances are generated from reference human translations on the TL side, so the MT output may be longer: it contains complete sentences rather than the immediate context of specific MWEs. In this case, we either use Recall-oriented metrics – e.g., WNM (Babych and Hartley 2004) – or, for Precision-oriented metrics, we swap the test and the reference files so, that the MT output becomes a reference.
4. We generate the evaluation results in the form of tables, where particular MWEs are ranked by BLEU or other automated scores. The resulting tables can be used by MT developers similarly to traditional risk-analysis tables: they can focus on highly-probable (i.e., most frequent) lexical errors with the greatest impact on quality (i.e., lowest BLEU/WNM for the concordance).

3. Experiment

We carried out an experiment for discovering mistranslated MWEs in the DARPA-94 MT evaluation corpus that contains approx. 35k words; it includes two independent human translations into English of 100 French news texts, as well as the output of 4 MT systems scored by human judges for *adequacy*, *fluency* and *informativeness*. Since there is no access to those systems, we used scenario 2(b) described above and generated MWEs on the Target (English) side. The resulting lists indicate, therefore, which TL MWEs were not properly generated by MT; nevertheless, in many cases it is possible to trace the discovered errors to dictionary gaps on the SL side.

We generated lists of continuous MWEs in a window of up to 5 words, and applied the *idf* filter (*log(idf) > 1*) to the edges of MWEs and the frequency filter (*freq(MWE) > 4*). For continuous MWEs a lower frequency filter can also yield good results, e.g., *freq(MWE) >1* (Sharoff et al., 2006). However, in our experiment MWEs also select concordances that are used for computing BLEU scores. Therefore, a higher threshold was chosen to enhance the reliability of the automated scores for concordances.

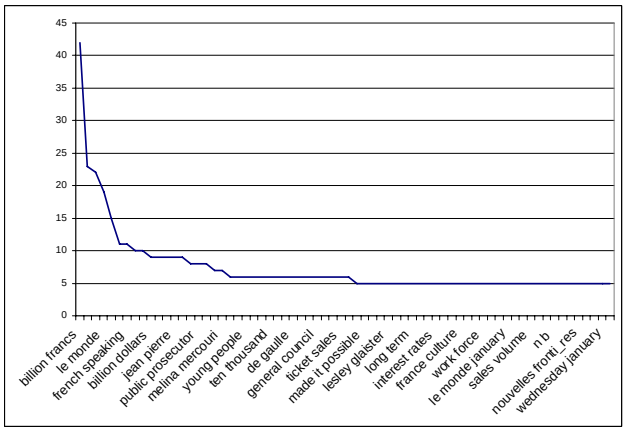

Chart 1. Frequency distribution of MWEs

MWEs were generated for the human *reference* translation. In total, 68 continuous MWEs passed both filters. Most typically these were 2-word MWEs, but there are also several 3- and 4-word MWEs. Frequencies of MWEs were in the range of 42 to 5. Chart 1 shows the corresponding frequency distribution.

The number of extracted MWEs is relatively small, so there is a need to use larger corpora in order to include more MWEs in the process of error analysis. Still, our approach can be applied to individual words, which have higher frequencies in smaller corpora. Using

discontinuous MWEs should also increase the number of constructions examined.

For each extracted MWE we generated aligned concordances. The concordance for the human *reference* translation (which was used as a *test* file for BLEU) contained at least 5 lines, each line including the MWE itself and up to 4 words to the left and to the right. Each of these lines was aligned with a full segment (typically – a paragraph) generated by the 4 MT systems and by another 'expert' human translator. In our experiment these segments were used as *test* files for the BLEU script. Note that we swapped *test* and *reference* files compared to the standard set-up of BLEU (where MT output is used as *test* and human translations – as a *reference*). The reason is that our experiment aims at discovering MWEs in human translations that were not properly generated by MT systems. BLEU is a *precision-oriented* metric, so it tests the absence of spurious N-grams in the test set. Therefore, in our experiment these interesting MWEs in human translations are moved to the *test* file and if they are not matched by anything in the MT output, they are treated as 'spurious' and the segment is penalised by a lower score.

We computed BLEU scores for each of our 68 concordances (using 1 reference and N-gram size up to 4) for each of the 4 MT systems and for the 'expert' human translation. Table 1 presents the scores for some interesting MWEs for each MT system and for the human translation. The MWEs are sorted by the BLEU score for Systran ('syst').

For MT output, low scores for the concordance of an MWE mean that it is not generated properly by the particular MT system. So we suggest that the highlighted MWEs are problematic for Systran and require developers' attention. The threshold is set at the system's average BLEU score of 2.7, which also coincides with a jump in the series of values.

Note that average scores can characterise general performance of an MT system, e.g., scores for human translation are higher than for MT output. Still, these scores are computed in a very different way than standard BLEU evaluation and correlation of the average with human judgements is lower than the figures reported for BLEU, which are in the region of 0.98 (Babych and Hartley, 2004). Still, these scores show high positive correlation with *adequacy*, and a slightly lower correlation with fluency, despite the fact that the corpus size is much smaller. Table 2 shows these correlation figures.

We checked contexts for some of the expressions in Table 1 in order to find out whether lower BLEU scores are due to sporadic mismatches (since the size of the evaluation sub-corpus in this case is much smaller than for standard BLEU evaluation), or whether lower scores indeed correspond to translation problems for these particular MWEs. In the majority of cases lower BLEU scores indeed correspond to consistently less fluent translations or mistranslations. Tables 3 and 4 illustrate such cases by comparing concordances for the human reference translation and MT output.

	Hum (exp)	**cand**	**glbl**	**ms**	**rev**	**syst**
credit lyonnais	*0.33*	0.16	0.16	0.1	0.12	0.1
work force	*0.37*	0.35	0.1	0.1	0.12	0.11
ticket sales	*0.26*	0.24	0.09	0.11	0.2	0.11
once again	*0.12*	0.09	0.09	0.15	0.09	0.11
french speaking	*0.48*	0.11	0.15	0.23	0.26	0.12
sales volume	*0.18*	0.13	0.1	0.11	0.11	0.12
public prosecutor	*0.21*	0.17	0.16	0.12	0.3	0.18
take place	*0.32*	0.17	0.14	0.15	0.34	0.18
term rates	*0.37*	0.25	0.12	0.2	0.35	0.19
press release	*0.23*	0.22	0.19	0.15	0.17	0.19
daily life	*0.39*	0.17	0.23	0.17	0.45	0.2
so-called	*0.38*	0.2	0.15	0.19	0.16	0.21
young people	*0.32*	0.1	0.1	0.18	0.16	0.28
managing director	*0.42*	0.22	0.19	0.42	0.21	0.31
minister of foreign affairs	*0.63*	0.59	0.29	0.54	0.18	0.33
examining magistrate	*0.36*	0.13	0.14	0.29	0.25	0.34
media library	*0.5*	0.17	0.11	0.16	0.32	0.34
other hand	*0.37*	0.16	0.66	0.46	0.63	0.39
prime minister	*0.54*	0.33	0.44	0.24	0.44	0.39
interest rates	*0.7*	0.39	0.2	0.44	0.52	0.41
made it possible	*0.23*	0.21	0.1	0.11	0.18	0.41
european union	*0.44*	0.33	0.45	0.5	0.46	0.45
general council	*0.43*	0.21	0.49	0.45	0.48	0.48
united states	*0.56*	0.28	0.41	0.35	0.53	0.62

…

	Hum (exp)	**cand**	**glbl**	**ms**	**rev**	**syst**
Average	*0.38*	0.22	0.22	0.25	0.29	0.27

Table 1 BLEU scores for MWEs

	r correl
Adequacy	*0.883*
Fluency	*0.620*
Informativeness	*0.380*

Table 2. Correlation of Average for all MWEs

*Fr:… Depuis le début du siècle, ses **effectifs** sont passés de 15000 à 2500 emplois…*	
Ref human	*Systran*
its **work force** has fallen from	its **manpower** passed from
believes that reducing the **work force** would	estimates that to touch **manpower** would
continues to reduce its **work force** in Europe	continues the reduction of its **manpower** in Europe
reducing its **work force** from	bringing back its **manpower** in

Table 3. MWE *work force*

*Fr:… Soit 53 % des **entrées** avec 40 % des écrans… La famille-fantôme fait mieux que la famille saint-bernard avec, respectivement, 75 000 (près de 160 000 en quinze jours) et 67 000 **entrées** (200 000 en trois semaines).*	
Ref human	*Systran*
this would be 53% of **ticket sales** with 40% of the screens	That is to say 53% of the **entries** with 40% of the screens
and 67 000 **ticket sales** (200 000 in three weeks	and 67 000 **entries** (200 000 in three weeks
with another 43,000 **ticket sales** during its fifth week	with 43 more 000 **entries** in fifth week

Table 4. MWE *ticket sales*

It can be seen from the tables that the MWEs were consistently translated less adequately compared to human translation. However, for MWEs with higher BLEU scores this was not the case: their translation was still adequate. Table 5 illustrates this for the MWE *minister of foreign affairs*, which is above the threshold of BLEU = 0.2.7.

*Fr:… Les négociations actuelles, patronnées par les Etats-Unis, sont menées par le **ministre** croate **des affaires étrangères**, Mate Granic, et le premier ministre bosniaque, Haris Silajdzic*	
Ref human	*Systran*
in paris the **minister of foreign** affairs stated friday	In Paris, the **Foreign Minister** declared, Friday
the israeli **minister of foreign** affairs Shimon Peres thought	the Israeli **Foreign Minister** Shimon Peres estimated
led by the **croat minister of foreign affairs** Mate Granic	carried out by the Croatian **Minister for the Foreign Affairs**, Mate Granic
the nigerian **minister of foreign affairs** babangana kingibe left	**The minister** Nigerian **of the Foreign Affairs**, Babangana Kingibe, flied away

Table 5. MWE *minister of foreign affairs*

These results are surprising, given the fact that BLEU is generally used only at 'higher' levels of evaluation: it offers high correlation with human judgements only at the level of an entire corpus, but not for individual texts or sentences. But it now appears that these scores have an additional '*island of stability*' at the level of individual lexicogrammatic constructions. Concordance-based evaluation provides a sufficiently focussed approach for these constructions, where BLEU scores become meaningful also at the micro-level. A possible explanation for this can be that the sub-corpus for evaluation of MWEs is collected in a very controlled way, which limits the noise factor.

4. Normalisation for translation variation

As we noted earlier, for MT output low BLEU scores for the concordance of an MWE mean that the MWE is not generated properly. However, we included a second human translation – the 'expert' translation – in our evaluation set, and for this human translation the meaning of lower BLEU scores is very different. If we suppose that professional human translators cannot be wrong very frequently, then lower scores for a given MWE mean that there are other legitimate ways to express the intended meaning. Therefore, generating that specific MWE is not essential for the content. Such expressions typically belong to the general lexicon and can be freely re-phrased in the same context. On the other hand, if a given MWE has a high BLEU score, then it was consistently inserted into the text by both human translators. Thus, it is more stable and possibly even obligatory for such contexts. Such expression are usually terms or other stable constructions which require specific and invariable translation equivalents.

Table 6 presents MWEs sorted by the BLEU scores for the 'expert' human translation. (Highlighting of problematic expressions for Systran is preserved, as in Table 1.)

	Hum (exp)	**cand**	**glbl**	**ms**	**rev**	**syst**
once again	*0.12*	0.09	0.09	0.15	0.09	0.11
sales volume	*0.18*	0.13	0.1	0.11	0.11	0.12
public prosecutor	*0.21*	0.17	0.16	0.12	0.3	0.18
press release	*0.23*	0.22	0.19	0.15	0.17	0.19
made it possible	*0.23*	0.21	0.1	0.11	0.18	0.41
ticket sales	*0.26*	0.24	0.09	0.11	0.2	0.11
take place	*0.32*	0.17	0.14	0.15	0.34	0.18
young people	*0.32*	0.1	0.1	0.18	0.16	0.28
credit lyonnais	*0.33*	0.16	0.16	0.1	0.12	0.1
examining magistrate	*0.36*	0.13	0.14	0.29	0.25	0.34
work force	*0.37*	0.35	0.1	0.1	0.12	0.11
term rates	*0.37*	0.25	0.12	0.2	0.35	0.19
other hand	*0.37*	0.16	0.66	0.46	0.63	0.39
so-called	*0.38*	0.2	0.15	0.19	0.16	0.21
daily life	*0.39*	0.17	0.23	0.17	0.45	0.2
managing director	*0.42*	0.22	0.19	0.42	0.21	0.31
general council	*0.43*	0.21	0.49	0.45	0.48	0.48
european union	*0.44*	0.33	0.45	0.5	0.46	0.45
french speaking	*0.48*	0.11	0.15	0.23	0.26	0.12
media library	*0.5*	0.17	0.11	0.16	0.32	0.34
prime minister	*0.54*	0.33	0.44	0.24	0.44	0.39
united states	*0.56*	0.28	0.41	0.35	0.53	0.62
minister of foreign affairs	*0.63*	0.59	0.29	0.54	0.18	0.33
interest rates	*0.7*	0.39	0.2	0.44	0.52	0.41

Table 6. MWEs sorted by 'expert' human BLEU

It can be seen from the table that general language expressions with greater contextual variability are at the top, while more stable terminological units are at the bottom.

This finding suggests that MT systems should be rewarded for having higher BLEU scores for more stable constructions, while being allowed greater freedom to deviate from less stable equivalents. Therefore, we take into account not only absolute values of BLEU for a

given construction, but also how different the score for an MT system is from the corresponding BLEU score for a human translation. In the general case BLEU for MT and for human translations are independent, but the measure of MT quality is how well they go together: whether we can reliably predict the difference between the MT and human scores given the raw MT score. Chart 2 illustrates this point. The horizontal axis shows values for *human* translation, and the vertical axis shows values for Systran.

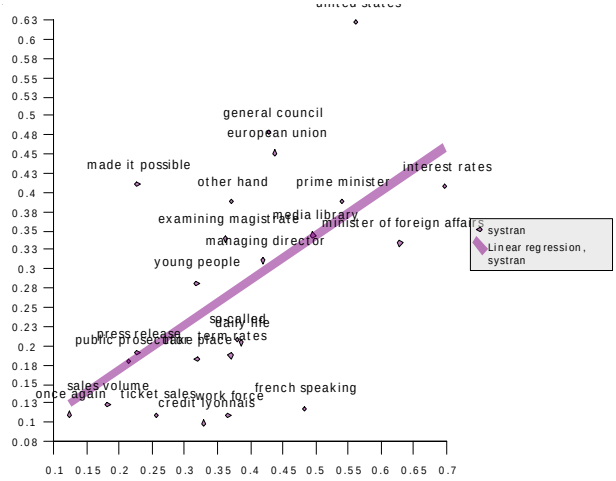

Chart 2. BLEU for human translation vs Systran MT

MWEs in this chart are located along 2 dimensions: MWEs closer to the right are more stable (more terminological), while those closer to the left belong to the general lexicon and can be more frequently rephrased. On the other hand, MWEs at the top are less problematic for Systran MT, and those at the bottom are more difficult. In an ideal case the points of the chart should be close to the diagonal line. Deviations from this line mean either that an MT output matches the human translation of a variable term (e.g., MWE *made it possible* in the top-left corner of Chart 2), or that it does not cover specific stable terms (e.g., MWE *French-speaking* in the bottom-right corner of Chart 2 – there is a gap in Systran's dictionary: …*of the Flemish francophonie*… instead of … *of the flemish french speaking community.*…).

We suggest that we can measure certain aspects of MT quality by the degree of agreement between BLEU scores for human translation and for MT. Such agreement can be captured by the correlation coefficient *r*. We compute it between two arrays of scores: the array of raw BLEU figures for an MT system, and the array of differences between these scores and BLEU for the human translation (for corresponding MWEs):

$$N = r\frac{\{bleu_{MT}(MWE)\}}{\{bleu_{MT}(MWE) - bleu_{humanTr}(MWE)\}}$$

We found that there is a high correlation between human judgements for *informativeness* and the N (normalised variation) score. Table 7 illustrates the correlation between N and each of the human evaluation parameters available for the DARPA corpus.

	cand	**glbl**	**ms**	**rev**	**syst**	**r corr with N**
N-score	0.13	0.38	0.25	0.45	0.38	
[ade]	0.68	0.71	0.71		0.79	0.72
[flu]	0.45	0.38	0.38		0.50	-0.02
[inf]	0.64	0.75	0.66		0.76	**0.97**

Table 7. Correlation: N-*score* vs human scores

The table suggests that for better, more *informative* MT systems there is a better agreement between BLEU scores for MT and the difference {MT *vs* human}: if BLEU is low, then the difference should be also low, which means that the human score is low as well. Thus MT is allowed to have low scores only for re-phrasable, highly variable expressions from the general lexicon.

5. Applications

The proposed approach can be useful in two main ways, without the need for human scores. Firstly, it can discover MWEs on the SL side or on the TL side which are, respectively, poorly translated by one or several MT systems, or not properly generated. Along this dimension our method is useful for MT developers in their efforts to discover the most typical lexical errors and improve the quality of their systems. It is equally useful for MT users who wish to extend their dictionaries before launching production in a new subject domain.

Secondly, our approach can also highlight MWEs which are usually translated correctly by MT systems. This information can be useful in the specification of Controlled Language or of MT-tractability using large-scale corpus data, and based on the performance of a particular state-of-the-art MT system.

We have shown that the N-score, which is a correlation coefficient between standard and normalised BLEU scores for individual MWEs, is a good predictor of human judgements about *informativeness* at the corpus level. Previously no automated metrics could approximate this quality parameter.

6. Future work

Future work will involve testing two commercial rule-based MT systems – Systran 6.0 and ProMT 8.0 – and Google's statistical MT system available on-line. We will generate translations for the French-English and (for the Euronews corpus) Russian-English directions. Current challenges for the project include: determining an optimal size of immediate context for the concordances; selecting the most revealing automatic metrics; (meta-)evaluation of the approach, e.g., using corpus-level human scores; and determining classes of MT error which most influence human evaluation scores.

7. References

Babych, Bogdan and Hartley, Anthony. 2004. Extending the BLEU MT Evaluation Method with Frequency Weightings. In *ACL 2004: the 42nd Annual Meeting of the Association for Computational Linguistics*.

Babych, Bogdan and Sharoff, Serge and Hartley, Anthony and Mudraya, Olga. 2007a. Assisting Translators in Indirect Lexical Transfer. *In ACL 2007*.

Babych, Bogdan and Hartley, Anthony and Sharoff,

Serge. 2007b. Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Machine Translation Summit XI*, 10-14 September 2007, Copenhagen, Denmark.

Estrella, Paula and Hamon, Olivier and Popescu-Belis, Andrei. 2007. How Much Data is Needed for Reliable MT Evaluation? Using Bootstrapping to Study Human and Automatic Metrics. In *Machine Translation Summit XI*, 10-14 September 2007, Copenhagen, Denmark.

Miller, Keith J. and Vanni, Michelle. 2005. Inter-rater agreement measures, and the refinement of metrics in the PLATO MT evaluation paradigm. In *Machine Translation Summit XI*., Phuket, Thailand, September 13-15.

Sharoff, S., Babych, B., Hartley, A. 2006. Using comparable corpora to solve problems difficult for human translators. In *Proceedings of COLING/ACL 2006 Conference*, Sydney.

Thurmair, G. 2007. Automatic evaluation in MT system production. In *Machine Translation Summit XI workshop: Automatic Procedures in MT Evaluation*, 11 September 2007, Copenhagen, Denmark.

White J. and O'Connell T. and O'Mara F. 1994. The ARPA MT evaluation methodologies: evolution lessons and future approaches. In *1st Conference of the Association for Machine Translation in the Americas*.

Reference-based vs. Task-based Evaluation of Human Language Technology

Andrei Popescu-Belis

IDIAP Research Institute
Centre du Parc – BP 592
CH-1920 Martigny, Switzerland
andrei.popescu-belis@idiap.ch

Abstract

This paper starts from the ISO distinction of three types of evaluation procedures – internal, external and in use – and proposes to match these types to the three types of human language technology (HLT) systems: analysis, generation, and interactive. The paper first explains why internal evaluation is not suitable to measure the qualities of HLT systems, and shows that reference-based external evaluation is best adapted to analysis systems, task-based evaluation to interactive systems, while generation systems can be subject to both types of evaluation. Some limits of reference-based external evaluation are analyzed in the case of generation systems. Finally, the paper shows that contextual evaluation, as illustrated by the FEMTI framework for MT evaluation, is an effective method for getting reference-based evaluation closer to the point of view of the users of a system.

1. Introduction

The nature of the evaluation methods that can be applied to human language technology (HLT) systems depends on the type of such systems, and more specifically on the place of language among their inputs and outputs. This paper considers the three types of evaluation synthesized in the ISO/IEC 9126 and 14598 standards – internal, external, and in use – and attempts to match them to an I/O-based typology of HLT – as analysis, generation or interactive systems. More specifically, we argue that: (1) analysis systems, which have language as their input, are best evaluated against manually built ground-truth samples of output; (2) interactive systems, which deal with series of linguistic input and output pairs, are best evaluated through their use by human subjects; and (3) generation systems, which produce linguistic output without human interaction, can be evaluated both ways, but with serious challenges in each case.

We describe first the ISO/IEC typology of evaluation types (Section 2) and our typology of HLT systems (Section 3), and outline the matching between the two (Section 4), showing that internal evaluation cannot significantly capture any of the qualities of HLT systems. Then, we argue that analysis systems are naturally submitted to reference-based external evaluation (Section 5), while for generation systems, reference-based and task-based evaluation have respective advantages and drawbacks, mainly a trade off between informativeness and cost (Section 6). We also pinpoint the potential risk of training a system for higher scores on a specific metric, regardless of its overall quality (Section 7). For interactive systems, the only feasible evaluation appears to be the task-based one, which can be carried out in more or less realistic settings (Section 8). Finally, we argue that adapting reference-based evaluation to the intended context of use of a system – as in the FEMTI guidelines for context-based MT evaluation – is a way to get the results of reference-based evaluation closer to the conclusions of task-based evaluation, and for a smaller cost (Section 9).

2. Three Types of Evaluation

The ISO/IEC standards for software evaluation, under the 9126 and 14598 series and then the SQuaRE framework (Azuma, 2001), have defined *software quality* as the "features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs" (ISO/IEC, 2001 : p. 11).

2.1. ISO/IEC Quality: Internal, External, in Use

According to ISO/IEC 14598-1 (1999 : p. 12, fig. 4) the software life cycle starts with an analysis of the user needs, determining a set of *external quality requirements*, which are then transformed into *internal* ones during the development phase. Once a system is implemented, it becomes possible to assess its *internal quality*, which is defined as "the totality of attributes of a product that determine its ability to satisfy stated and implied needs" (ISO/IEC, 1999, §4.15). Internal quality is assessed without running the system – though not necessarily "internally" by its developers – by measuring internal parameters that are known to have an impact on quality, for instance the size of a dictionary, the number of rules, etc. Because for HLT systems the contribution of such attributes to perceived quality cannot be taken for granted, internal measures are seldom used in HLT, though they are sometimes used for advertising products[1].

External quality is evaluated by running the system and applying external measures, which are "indirect measure[s] of a product derived from measures of the behaviour of the system of which it is a part" (ISO/IEC, 1999, §4.6). Finally, *quality in use* is the extent to which a system really helps users fulfil their tasks (ISO/IEC, 2001: p. 11). To summarize, in ISO terms:

> *Internal metrics measure the software itself, external metrics measure the behavior of the*

[1] Dictionary size is often announced by vendors of translation software, e.g. Linguatec ("more than 3.8 million entries"), Systran ("millions of words and expressions"), or Word Magic ("900,000 uninflected entries").

computer-based system that includes the software, and quality in use metrics measure the effects of using the software in a specific context of use (ISO/IEC, 2001).

Quality in use is often decomposed in terms of effectiveness, efficiency, user satisfaction and safety (ISO/IEC, 2004) while internal and external qualities belong in six categories: functionality, reliability, usability, efficiency, maintainability and portability. In many cases of HLT evaluation, it is the qualities under functionality that are the focus of evaluation.

According to ISO/IEC, quality in use does not follow automatically from external quality, as it is not possible to predict all the results of using the software before it is operational in its intended context of use. In what follows, we will mainly use the important distinction between external evaluation – often based on the comparison with ground truth output – and evaluation in use, which we now compare to other distinctions.

2.2. Relation to Other Types of Evaluation

The HLT evaluation community often opposes black-box to glass-box evaluation, which correspond roughly to the external vs. internal ISO/IEC types. The only unclear case is when the internal parameters of a system are examined during its execution, which would probably qualify as external *and* glass-box.

Sparck Jones and Galliers (1996, p.19) oppose intrinsic evaluation ("relating to a system's objective") to extrinsic evaluation ("relating to its function i.e. to its role in relation to its setup purpose"). In ISO/IEC terms, this corresponds to the distinction between external evaluation and evaluation in use: despite the potentially misleading analogy, 'external' evaluation remains 'intrinsic' (it looks at a system's own performance), while 'evaluation in use' is 'extrinsic' (it looks at a system's utility in a given setup).

3. An I/O-based Typology of HLT Systems

In order to study the most adapted evaluation techniques for HLT systems, we propose to classify them according to the occurrence of language in their data: in the input to a system, in its output, or in both. Additionally, the system may or may not require an interaction with a human user in order to produce its global results.

Type A systems ('A' stands for analysis or annotation) have language as an input only, and they often they perform classification of the linguistic material into a small number of categories; examples of type A tasks are POS tagging, WSD, or reference resolution. Type G systems ('G' is for generation) have language only as an output, for instance when generating weather reports from non-linguistic data. Type AG systems have both linguistic input *and* output, including tasks such as machine translation, automatic summarization, or question answering. Finally, type I systems, or more accurately type AGI, are language-based human-computer dialogue systems.

This classification appears to be exhaustive and non-ambiguous, as shown elsewhere (Popescu-Belis, 2008) by analyzing the HLT domains and applications from two encyclopaedias of HLT and NLP (Dale, Moisl & Somers, 2000; Mitkov, 2003).

The results expected from a type A system can generally be defined by a unique ground truth or gold standard annotation, possibly accompanied by an estimate of its reliability, if human judges agree less than perfectly upon this gold standard. In the case of G or AG systems, it is however impossible to find a unique gold standard, or to enumerate all acceptable results, due to the variability of natural language. In this case, it is still possible to provide a sample of the set of acceptable results, produced by human subjects. Alternatively, given the output of a G or AG system, a human judge can decide whether this output belongs or not to the ground truth, i.e. whether it is a perfect answer or not.

4. Matching Types of Evaluation with Types of Systems

Following the definitions above, the main point of this paper is to discuss whether some of the three ISO-based types of evaluation are better suited to some of the types of HLT systems. In principle, according to ISO, all types of HLT systems can (and should) undergo all types of evaluation, at the corresponding stages of their development lifecycles. However, this is clearly not feasible in the HLT research community. More precisely, we will argue in the next sections that the following rules characterize best practice in HLT evaluation:

- *internal* evaluation is seldom of interest: it is not enough informative for HLT systems, as it cannot predict external qualities or qualities in use;
- for type A systems, *external* evaluation using ground-truth data is informative and cost-effective;
- for type G and AG systems, there is a trade-off in informativeness vs. cost when switching from reference-based *external* evaluation to evaluation *in use* or *task-based*;
- for type I systems, only evaluation *in use* is informative enough, but can be performed in more or less realistic conditions.

The first point is justified by the observation that the behaviour of very few HLT systems can be completely predicted from their internal properties, unlike more deterministic software. Linguistic problem-solving is most often based on heuristics that show no clear relation between internal properties and external performance: e.g., for a parser, the number of syntactic rules is only marginally correlated with parsing accuracy or coverage. Of course, some generic qualities such as portability can be measured internally, but such qualities are seldom the focus of HLT evaluation, which generally aims at on functionality, i.e. the capacity to perform an intended linguistic function. In addition to aspects of functionality, speed is sometimes taken into account as well, but again it is generally not measured using internal metrics.

5. Evaluation of Type A Systems: Reference-based External Metrics

The linguistic functionality of annotation or analysis systems is most often measured by comparing their results to ground truth annotations produced by human judges. Such reference-based external metrics are generally expressed as (pseudo)distances between a system's response on some test data and the expected response or set of responses, as defined by human judges, and are

generally computed automatically. Whether or not the set can be determined with enough precision is a problem related not to HLT, but to the study of the respective linguistic capacity in human subjects.
This of course does not exclude evaluation in use – in case a specific use of the annotations was identified – but, in most cases, the results of reference-based evaluation are good indicators of performance in use, while being considerably cheaper to obtain, and more reliable in the sense that the measures can be repeated at will, with the same results on the same data.

6. Type G/AG Systems: Reference-based vs. Task-based Evaluation

For HLT systems that generate linguistic output (type G or AG), reference-based evaluation can only be applied if one can determine a distance between the system's response and a set of ground truth responses. The problem is that this set is potentially very large, has fuzzy borders, and is generally known only through a small number of samples that are collected from human subjects. The quality of a system's output, i.e. a distance to the set of acceptable responses, can either be judged directly by human evaluators, using or not the samples of acceptable responses, but it can also be inferred automatically from the distance to the samples. Therefore, while for type A systems the human judges define explicitly the set of acceptable responses, for type G/AG systems they merely verify mentally, using their linguistic competencies, whether a response belongs or not to the set (which is vastly larger for G/AG systems than for A ones).
The design of reference-based automatic metrics for type G/AG systems has been formulated as a training problem (Soricut & Brill, 2004), which can be solved using machine learning. The distance to the samples, and its average when several samples are available, are often adjusted over training data to match human judgments of quality.
A typical example are machine translation (MT) systems, for which the BLEU metric (Papineni, Roukos, Ward & Zhu, 2001) estimates the quality of automatically translated sentences based on their similarity to up to four human-translated versions of the same source sentence (BLEU was manually optimized to match human judgments of adequacy and fluency). The limits of reference-based evaluation metrics for MT have been widely discussed (Culy & Riehemann, 2003; Callison-Burch, Osborne & Koehn, 2006), but the cost-effectiveness of these methods compensates their inadequacy to human judges in many cases. As MT quality gets closer to human translators, the defects of reference-based metrics become more obvious (Popescu-Belis, 2003).
Task-based evaluation is the other option for assessing the quality of G/AG systems. This method appears to be more informative than reference-based evaluation as it measures "directly" the satisfaction of user needs (which is the very definition of quality in ISO terms) and considers all the quality aspects of a system, but comes at a significantly higher cost, as each measurement involves a large number of human subjects. Also, as each measurement has to be repeated when the system changes, task-based evaluation is much less generic than reference-based evaluation.
Turning again towards MT evaluation as a case study, task-based evaluation was discussed by (White, Doyon & Talbott, 2000) among others, and has recently inspired a metric named HTER, which estimates the utility of MT output based on the human post-editing effort required to correct it (Snover, Dorr, Schwartz, Micciulla & Makhoul, 2006; Przybocki, Sanders & Le, 2006).

7. A Risk of Reference-based Evaluation for Type G/AG Systems

As we have shown, reference-based metrics approximate the quality of the output from its "distance" to a small number of samples of desired output. When evaluators define such approximations in order to measure as accurately as possible output quality, providing data and software to compute the distances, these (pseudo)metrics are soon used by developers to improve their systems. Therefore, the metrics start being incorporated into the optimization criteria of the systems, especially those based on machine learning approaches. Hence, two potential problems may arise.
Firstly, if the metric is quite imperfect, training a system to improve its scores will not improve its true quality, as it can be assessed by independent metrics. Secondly, although developers are not allowed to train their systems on test data (a fact that would invalidate the evaluation results), they can train the system to obtain higher scores for a given evaluation metric, regardless of the training/test data. This becomes a problem when the metric is poorly matched to users' needs, and is more acute as the system gets higher scores for the given metric.
A simple fix to both problems, still within the framework of reference-based evaluation, is to use several evaluation metrics instead of one, and consider that only concordant variations of all metrics represent significant variations of output quality. Another, more radical approach would be to use a previously unseen metric for an official evaluation, although it is not likely that developers would accept such a challenge.
For instance, in the MT case study, BLEU scores are generally improved if MT output is "smoothed" using a language model, regardless of the resulting meaning. To avoid this kind of tuning to BLEU, a possible solution is to use several automated metrics, some of which are not n-gram based, as in the CESTA French evaluation campaign (Hamon, Popescu-Belis, Choukri, Dabbadie, Hartley, Mustafa El Hadi, Rajman & Timimi, 2006). The NIST TIDES campaigns in the USA also used internally several metrics, some automatic and some human (for validation), although only BLEU scores were reported finally (NIST, 2006).

8. Evaluation in Use for Interactive Systems

Type I systems (or AGI) do not produce directly a result based on input data, but require a series of interactions with a human user, in which language may appear in the input or output, and most often in both, as is the case with human-computer dialogue systems. Such systems have been called 'symbiotic' ones (King & Underwood, 2006), and the one-input-to-one-output view does not suit them: hence, reference-based evaluation metrics are difficult to

apply to such systems, due to the large variety of possible input/output combinations at each step of the interaction. Therefore, type I systems are mainly evaluated using task-based approaches or evaluation in use, requiring human subjects to interact with the system (Dybkjær, Bernsen & Minker, 2004; Bevan, 2001). The top level parameters that are evaluated are:

- effectiveness: is the task is accomplished or not?
- efficiency: is the task accomplished efficiently or quickly?
- user-satisfaction;
- safety: seldom measured for HLT systems[2].

The limits of this type of evaluation are its relatively higher cost with respect to reference-based evaluation, due to the use of human subjects, and the difficulty to generalize the obtained results to slightly different tasks or contexts of use.

The evaluation of interactive systems may use two slightly different approaches, depending on what level of generality is sought, and which human subjects are available. One can distinguish *task-based* evaluation from genuine *evaluation in use*, defining the first one as evaluation using an idealized setting and generic subjects (or even another software interacting with the first one), while the second one is the evaluation in the final, intended context of use, with a sample of the final users. Task-based evaluation can be applied to research prototypes, while evaluation in use seems reserved to end-user products.

Meeting browsers are a prototypical example of interactive systems: they allow search and browsing of large multimedia recordings of meetings in order to find information that is relevant to the human users. Initial experiments in the evaluation of meeting browsers have defined reusable resources and metrics for task-based evaluation, but have also shown the difficulty to reduce the variance of responses from human subjects (Popescu-Belis, Baudrion, Flynn & Wellner, 2008).

9. Context-based Evaluation: Between Reference-based and Evaluation in Use

Our analysis has tried to match two typologies, one for evaluation methods (internal, external, in use) and the other for HLT systems (A, G/AG, I). A question arising at this point is the following one: where does *contextual evaluation* – a recent trend in the evaluation of HLT systems – belong in our analysis? This trend is best exemplified by the FEMTI guidelines for MT evaluation (Hovy, King & Popescu-Belis, 2002; Estrella, Popescu-Belis & Underwood, 2005) which emphasize the influence of the intended context of use of a system on the evaluation metrics used to assess its quality, i.e. the need to define a *contextual quality model*.

We hypothesize that contextual evaluation such as the FEMTI guidelines might offer a promising compromise between reference-based and task-based approaches, when neither approach is optimal. On the one hand, the methods contained in FEMTI-style guidelines cover both reference-based and task-based evaluation metrics, but at least in the case of MT systems, there is a predominance of reference-based ones, related to external qualities (FEMTI's generic quality model is based on the ISO top level external qualities). Therefore, using quality models inspired from FEMTI is a cost-effective approach to evaluation, if reference data and metrics can be found for each quality attribute that is evaluated.

On the other hand, FEMTI argues that the set of evaluation metrics and their respective weights must be adapted to the intended context of use of the system, as shown within the EAGLES and ISLE projects (EAGLES Evaluation Working Group, 1996; Hovy, King & Popescu-Belis, 2002). This observation has inspired the FEMTI framework for MT evaluation but also user-based proposals for the evaluation of information retrieval systems (Sparck Jones, 2001; Chaudiron, 2004). These are all significant steps towards considering the role of human users of a system when defining an evaluation. The goal of FEMTI is thus to generate evaluation plans that grasp the qualities of a system as close as possible to task-based evaluation, but without the high costs and reduced generality of this type of evaluation.

10. Conclusion

This paper has discussed the relationship between various types of evaluation and various types of HLT systems. While annotation systems can be evaluated using mostly reference-based metrics and interactive systems must be evaluated using task-based approaches, generation systems are more challenging, as neither reference-based, nor task-based methods offer a satisfactory compromise between the cost of an evaluation (and hence its reproducibility) and its informativeness (the capacity to find the "real" qualities of a system). In this case, contextual evaluation exemplified by the FEMTI guidelines offers a principled way to use a set of reference-based metrics that is adapted to the intended tasks and users of a system.

11. Acknowledgments

The author acknowledges the support of the Swiss National Science Foundation's grants n. 200021-103318 and 200020-113604 for MT evaluation projects, and of the IM2 National Center of Competence in Research.

12. References

Azuma M. (2001). SQuaRE: The Next Generation of the ISO/IEC 9126 and 14598 International Standards Series on Software Product Quality. *Proceedings of Escom 2001 (12th European Software Control and Metrics Conference)*, London, UK, pp. 337-346.

Bevan N. (2001). International Standards for HCI and Usability. *International Journal of Human-Computer Studies*, vol. 55, pp. 533-552.

Callison-Burch C., Osborne M. and Koehn P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. *Proceedings of EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, Trento, Italy, pp. 249-256.

[2] An apocryphal example of safety evaluation (or lack thereof) is the proposal for MT known as "helicopters in Vietnam", which suggested to evaluate MT of technical documents (here, for helicopter maintenance) by the number of failures of the equipments that were repaired using translated documents.

Chaudiron S. (2004). La place de l'usager dans l'évaluation des systèmes de recherche d'informations. In S. Chaudiron (ed.), *Évaluation des systèmes de traitement de l'information*, Paris, Hermès, pp. 287-310.

Culy C. and Riehemann S. Z. (2003). The Limits of N-Gram Translation Evaluation Metrics. *Proceedings of Machine Translation Summit IX*, New Orleans, Louisiana, USA, pp. 71-78.

Dale R., Moisl H. and Somers H. (ed.) (2000). *Handbook of Natural Language Processing*. New York, NY, USA, Marcel Dekker.

Dybkjær L., Bernsen N. O. and Minker W. (2004). Evaluation and Usability of Multimodal Spoken Language Dialogue Systems. *Speech Communication*, vol. 43, n. 1-2, pp. 33-54.

EAGLES Evaluation Working Group (1996). *EAGLES Evaluation of Natural Language Processing Systems*. Final Report Center for Sprogteknologi, EAG-EWG-PR.2 (ISBN 87-90708-00-8).

Estrella P., Popescu-Belis A. and Underwood N. (2005). Finding the System that Suits you Best: Towards the Normalization of MT Evaluation. *Proceedings of 27th ASLIB International Conference on Translating and the Computer*, London, UK, pp. 23-34.

Hamon O., Popescu-Belis A., Choukri K., Dabbadie M., Hartley A., Mustafa El Hadi W., Rajman M. and Timimi I. (2006). CESTA: First Conclusions of the Technolangue MT Evaluation Campaign. *Proceedings of LREC 2006 (5th International Conference on Language Resources and Evaluation)*, Genova, Italy, pp. 179-184.

Hovy E. H., King M. and Popescu-Belis A. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, vol. 17, n. 1, p.1-33.

ISO/IEC (1999). *ISO/IEC 14598-1:1999 (E) -- Information Technology -- Software Product Evaluation -- Part 1: General Overview*, Geneva, International Organization for Standardization / International Electrotechnical Commission.

ISO/IEC (2001). *ISO/IEC 9126-1:2001 (E) -- Software Engineering -- Product Quality -- Part 1:Quality Model*, Geneva, International Organization for Standardization / International Electrotechnical Commission.

ISO/IEC (2004). *ISO/IEC TR 9126-4:2004 (E) -- Software Engineering -- Product Quality -- Part 3:Quality in Use Metrics*, Geneva, International Organization for Standardization/ International Electrotechnical Commission.

King M. and Underwood N. (2006). Evaluating Symbiotic Systems: the Challenge. *Proceedings of LREC 2006 (Fifth International Conference on Language Resources and Evaluation)*, Genoa, Italy, pp. 2482-2485.

Mitkov R. (ed.) (2003). *The Oxford handbook of computational linguistics*. Oxford, UK, Oxford University Press.

NIST (2006). *NIST 2006 MT Evaluation Official Results*. National Institute of Standards and Technology, http://www.nist.gov/speech/tests/mt/mt06eval_official_results.html.

Papineni K., Roukos S., Ward T. and Zhu W.-J. (2001). *BLEU: a Method for Automatic Evaluation of Machine Translation*. Research Report, Computer Science IBM Research Division, T.J.Watson Research Center, RC22176 (W0109-022).

Popescu-Belis A. (2003). An experiment in comparative evaluation: humans vs. computers. *Proceedings of Machine Translation Summit IX*, New Orleans, Louisiana, USA, pp. 307-314.

Popescu-Belis A. (2008). Le rôle des métriques d'évaluation dans le processus de recherche en TAL. *T.A.L. (Traitement Automatique de la Langue)*, vol. 48, n. 1, pp. 25.

Popescu-Belis A., Baudrion P., Flynn M. and Wellner P. (2008). Towards an Objective Test for Meeting Browsers: the BET4TQB Pilot Experiment. In A. Popescu-Belis, H. Bourlard et S. Renals (ed.), *Machine Learning for Multimodal Interaction IV*, Berlin, Springer-Verlag, pp. 108-119.

Przybocki M., Sanders G. and Le A. (2006). Edit Distance: A Metric for Machine Translation Evaluation. *Proceedings of LREC 2006 (5th International Conference on Language Resources and Evaluation)*, Genova, Italy, pp. 2038-2043.

Snover M., Dorr B., Schwartz R., Micciulla L. and Makhoul J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of AMTA 2006 (7th Conference of the Association for Machine Translation in the Americas)*, Cambridge, MA, USA.

Soricut R. and Brill E. (2004). A Unified Framework For Automatic Evaluation Using 4-Gram Co-occurrence Statistics. *Proceedings of ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics)*, Barcelona, Spain, pp. 613-620.

Sparck Jones K. and Galliers J.R. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*, Berlin / New York, Springer-Verlag.

Sparck Jones K. (2001). Automatic language and information processing: rethinking evaluation. *Natural Language Engineering*, vol. 7, n. 1, pp. 29-46.

White J. S., Doyon J. B. and Talbott S. W. (2000). Determining the Tolerance of Text-Handling Tasks for MT Output. *Proceedings of Second International Conference on Language Resources and Evaluation (LREC'2000)*, Athens, Greece, vol. 1, pp. 29-32.

FEIRI: Extending ISLE's FEMTI for the Evaluation of a Specialized Application in Information Retrieval

Keith J. Miller
The MITRE Corporation
7515 Colshire Dr., McLean, VA 22102
keith@mitre.org

Abstract

This is intended as an informal position paper discussing several of the issues raised by the organizers in the call for papers of this workshop on evaluation. In particular, I address some of the questions regarding the potential for the use of task-based evaluation for a technology in the IR realm — specifically, multicultural name matching and identity resolution. I consider possible extensions of the FEMTI framework for evaluation of identity matching technologies, and I briefly consider the applicability of automated or hybrid (human/automated) evaluation metrics to this technology area.

1. Introduction

1.1. Situation of the Paper

In this position paper, I discuss several of the issues raised by the organizers of this workshop on evaluation in the call for papers. In particular, I address some of the questions regarding the potential for the use of task-based evaluation for a technology in the IR realm – specifically, multicultural name matching and identity resolution. I consider possible extensions of the FEMTI (Hovy *et al*, 2002; see also **http://www.issco.unige.ch/femti**) framework for evaluation of identity matching technologies. For purposes of consistency, I refer to the port of FEMTI (Framework for the Evaluation of MT in ISLE) to the IR domain as "FEIRI" (Framework for the Evaluation of IR in ISLE). It should be noted that most of the comments in this paper focus on the quality of the multicultural name matching system's output – that is, the effectiveness of its matching – rather than on other features of the system (such as speed of throughput, for example). This is due both to the principal focus of my work and also to the focus of much of the work I was involved with in the Machine Translation (MT) evaluation aspects of ISLE/FEMTI. Finally, I briefly consider the applicability of automated or hybrid evaluation metrics to multicultural name matching. Note that this paper does presuppose some familiarity with FEMTI, and the work that led up to it. For background information of this type, see the references section as well as the URLs embedded throughout the paper.

1.2. Multicultural Name Matching

Given the increased mobility of people in our ever-shrinking world, we can no longer expect – if we ever could – to interact only with people from our own culture. In everyday interactions, we come into contact with individuals having cultures, beliefs, and customs that are different from our own.

One way in which world cultures differ is in their naming practices, that is in the conventions followed when assigning or using an individual's personal name. It is thus no more possible to ignore cultural diversity in our data processing practices than it is to ignore it when dealing with individuals face-to-face. But we do this when we force multicultural name data into Anglo-centric name models, such as the one depicted in Figure 1, which are incorrect for many cultures.

{Mr.|Ms.} + First Name + Middle Name + Last Name

Figure 1: An Anglo-centric Name Model

Simply asserting this, however, will not alter the myriad ways in which names are altered – one might say mangled – as they travel through data processing systems designed by people who are not cognizant of their specialized processing requirements. Thus, multiculturally-sensitive name matching systems must be able to compensate for this by retrieving names regardless of what violence may have been perpetrated upon them as they passed through the system, and they must also be tolerant of similarly damaged queries.

David R. Smith David Smythe D Smith	'Abd Al Rahman Abdurrahman
Maria Gonzalez Maria Gonzales Torres	William Morrow Guillaume Moreau

Table 1: Matching Variant Forms of Personal Names (note: names within the same cell are intended to match[2])

The ways in which personal naming practices differ between cultures lead to the many ways that these names may vary from their "original" forms[1] when we finally

[1] We consider here mainly names in roman script. The definition of the "original" romanized form of a name that did not originate in roman script is an issue that we will put aside for the present.

encounter them, be it in structured or semi-structured data, or in running text. These variation types will be discussed further below, but examples of some names that should likely match[2] even though they are not identical are found in Table 1.

2. Evaluation of Multicultural Name Matching

2.1 Potential for Use of Task-Based Evaluation for this Specialized IR Task

Different types of evaluation are called for at different points in the lifecycle of Human Language Technology (HLT) applications. Internal and diagnostic evaluations, including regression tests, are necessary during the research and development cycle. However, some time before deployment, all HLT applications should be subjected to some type of task-based evaluation. That is, it should be determined if the application in question can perform at a level necessary to support the intended end users of the application in the task that they are assigned to perform. Multilingual name search applications, as a special case of information retrieval applications, are no exception to this rule.

One might argue, however, that the leap from the application of some of the standard metrics in IR (that is precision, recall, and F-Score) to a task-based assessment of the IR systems in question is much shorter than that from a standard MT metric (if such a thing exists) to task-based MT evaluation. Some key differences between MT evaluation and evaluation of multicultural name search are summarized in Table 2, below.

	Machine Translation	Multicultural Name Matching
Aspects of output quality to measure	many	few
Number of metrics	many	few
Mapping between metrics and effect on tasks	less well-understood	better understood

Table 2: Comparison of Features of MT and MNM Evaluation

Whereas in MT evaluation, the mapping between the various aspects of the quality of a translation system's output (and the metrics that measure that quality) and the tasks that can be performed on that output is not clear, in multicultural name search, the aspects of output quality are limited (to whether a system produced the correct returns, and perhaps the ranking of those returns as the two most obvious), and the mapping to appropriate tasks or use cases is more a matter of tolerance for differing performance levels on the metric(s) that measure(s) these features. In task-based evaluation of both MT and MNM systems, a large amount of thought must go into the definition of the task; however, in order to choose the best candidate systems on which to perform a task-based MT evaluation, the evaluators must further give considerable thought to which quality characteristics of the output affect performance of that task, and choose the systems that perform best on metrics that measure those characteristics. Because of the relatively well-defined metrics for MNM evaluation as compared to MT evaluation, it may be that whereas the power of FEMTI lies in the mapping from use cases to quality characteristics to metrics, the utility of FEIRI may lie elsewhere. I will consider this possibility below[3].

2.2. Remarks on the Naturalness of a FEMTI-like Tool for Evaluation of Multicultural Name Search

Given that over the years we have found the EAGLES "7-Step Recipe" for evaluation [4] to be a useful and methodologically-sound way to organize and perform evaluations of HLT systems, when we were approached to evaluate multicultural name matching systems we began by consulting this resource. As such, we have already been conducting our evaluations of name search systems in the spirit of ISLE, beginning by taking into account the purpose of the evaluation and the use context in which the system is to be evaluated (see, for example (Arehart and Miller, 2008), (Miller *et al*., 2008), (Lloyd and Miller, 2007)).

It is not a coincidence that the FEMTI framework emerged in the natural progression of the EAGLES and ISLE work on MT evaluation. For IR, like for MT, beginning the evaluation with the above-mentioned mindset places the evaluator squarely in the first of the two FEMTI taxonomies – or, in the left-hand taxonomy of FEMTI as it is presented on the web[5]. This is the portion of FEMTI that helps the evaluator to define the use context for the systems to be evaluated. The natural inclination, then, is to start asking the questions that will cause the evaluator to move to the second (or right-hand) FEMTI taxonomy – the portion of FEMTI that considers system characteristics that are important to that task. Having chosen the system characteristics, it would naturally be helpful to have a system that could suggest evaluation metrics that are appropriate to measure those characteristics relative to the task set out in the first part of the taxonomy.

[2] for some use cases, at least

[3] This is not to claim that there is no ongoing research to be performed on IR metrics. On the contrary, interesting work is going on in this area. However, metrics being considered all seem to be variations revolving around a common theme, as opposed to the case in MT evalutaion, where the field still seems to be wide open.

[4] http://www.issco.unige.ch/projects/eagles/ewg99/7steps.html

[5] http://www.issco.unige.ch:8080/cocoon/femti/st-clasifFrame.html, pictured in Figure 3 at the end of this paper.

As alluded to in the previous section, however, when it comes to choosing metrics, we may find ourselves in a different state of affairs for IR as compared to MT. I will discuss this further in Section 3.

3. From FEMTI to FEIRI

3.1 Differences in Requirements for Evaluation of MT and IR: Evaluation Requirements

In this section, I consider the elements of the FEMTI taxonomy of evaluation requirements and their portability to a similar taxonomy in FEIRI.

3.1.1 FEMTI 1.1: Purpose of Evaluation

There are many elements of FEMTI that will carry over to almost any HLT evaluation – or to almost any system evaluation for that matter. The taxa under heading 1.1, which all name evaluations of different types, and describe their purposes, are certainly applicable to IR, and would also be applicable to most any HLT evaluation. Of course, the notes accompanying each taxon (relevant qualities – from part 2, and references in particular) would have to be updated to be applicable to IR. In addition, there are times when notes (as for taxon 1.1.4 "Operational Evaluation") or definition (as for taxon 1.1.3 "Declarative Evaluation" would also have to be modified, as they contain information that is specific to MT. In general, the entirety of FEMTI would have to be scrubbed for such MT-specific information in the annotations accompanying the more generally-applicable taxa.

3.1.2 FEMTI 1.2: Characteristics of the Translation Task

In FEIRI, the tasks listed under this taxon – "assimilation", "dissemination", and "communication" – would be replaced with IR-specific tasks. In the case of multilingual name matching, these tasks might include

- positive matching: searching for individuals that are supposed to be present in the data list, such that they may be accorded some benefit (e.g. access to an event, access to goods / money, etc.)
- negative matching: searching for individuals who should not be present in the data list. Such a list might include, for example, people from whom a business does not wish to accept checks as payment
- data cleansing: searching a data list against itself in order to remove duplicates that result from variant representations of the same name being added to the list

These are just several sample possible task categories for evaluation of multicultural name matching. Tasks like these may be interesting in that they map to various risk levels – for example the risk associated with missing a match in the "positive matching" scenario may be low (a person may be delayed in getting something to which they are entitled, resulting in an irritated person), whereas the risk associated with missing a record in the "negative matching" scenario may be much more serious (failing to retrieve the name of a patient from a list of all patients who are allergic to a certain medication may result in that patient receiving the medication, which could result in death). As suggested in section 2, it may be in this mapping from risk tolerances associated with task categories to evaluation elements in the second half of the taxonomy that FEIRI has its strength.

3.1.3 FEMTI 1.3: Input Characteristics (Author and Text)

Given the cross-cultural variation in naming practices, the closest FEIRI analog to FEMTI's 1.3.1, "Document Type" and its sub-taxa "Genre" and "Domain" is the cultural origin of the names being processed. Concerning 1.3.2, "Author Characteristics", as is the case with MT evaluation, FEIRI users may or may not have access to any information about the "author" of the names that are being processed. However, if this information is available, it would certainly be relevant to the evaluation. Additionally, it would be interesting to know if the names being processed originated from a primarily spoken or written source.

Taxon 1.3.3 "Characteristics related to sources of error" is quite interesting in relation to evaluation of name search. As stated above, it is primarily through the various errors that occur due to lack of sensitivity to the ways in which names differ between cultures that many of the problems in multicultural name matching arise. Because of this, in other work we have developed a taxonomy of name variation (Miller *et al*, 2008). This taxonomy, from which mappings to the taxa under 1.3.3 can be derived, can be seen in Figure 2.

As with the evaluation of MT, some of the errors may be "intentional" (as defined in FEMTI taxon 1.3.3.1), such as alternate spellings, transliterations, abbreviations, initials, nicknames, diminutives, and translation variants. Some may be "medium-related" (FEMTI taxon 1.3.3.2), to wit, we find OCR errors as well as fielding variation, and truncation, perhaps due to suboptimal data exchange mechanisms. Typos as well as some of the other variation types *may* fall under FEMTI taxon 1.3.3.3 "performance-related error sources". Although it may be difficult to determine intentionality or lack thereof in all cases (therefore making it impossible to determine whether errors were due to performance factors or intentional), a thoughtful cross-mapping between the error sources and taxa in the name variation taxonomy may be useful in planning MNM evaluations. Note that some errors, particularly in input of multicultural names, may be due not to performance, but to competence, a distinction that is alluded to by the inclusion of the author characteristics (FEMTI 1.3.2), but not carried through to the sources of error (FEMTI 1.3.3).

In FEIRI, it may be that it is exactly these sources of error that we map to the second part of the taxonomy – not

to quality characteristics in the output, but perhaps to characteristics of the test data on which the evaluation will be based.

Element Variations

- Data Errors
  - OCR
  - Truncation
  - Typo
- Particles
  - Particle Segmentation
  - With/Without Particle(s)
- Short Forms
  - Abbreviation
  - Initials
- Spelling
  - Alternate Spelling
  - Transliteration
- Nicknames and Diminutives
- Translation Variant
- Other Element Variation

Structural Variations

- Deletion/Addition
- Fielding Variation
- Permutation
- Placeholder for Missing Information
- Segmentation of Elements
- Other Structural Variation

Other Variations

- Alias/AKA
- Non-variant
- Undetermined

Figure 2: Name Variation Taxonomy

3.1.4 FEMTI 1.4: User Characteristics

The taxa under "User Characteristics," which in FEMTI have to do with the end user's linguistic proficiency map quite nicely to the domain of evaluation of multicultural name searching. Those referring to "source language" and "target language" could refer to "culture of query name" and "culture of matched record" (or "culture of data list record"), respectively. The user's familiarity with naming practices in those two cultures will greatly affect their ability to use the system, and thus the overall utility of the system to them. To this might be added another sister taxon, titled "mediating culture", which would account for artifacts introduced when a name comes to its current form by virtue of having passed through a language or culture other than the one in which it originates[6]. The taxa pertaining to the organizational user (under FEMTI 1.4.2) apply equally to users of MNM systems, with the requisite modifications to change the object of measurement from quantities and speed of translation to quantities and throughput of MNM queries.

3.2 Differences in Requirements for Evaluation of MT and IR: System Characteristics

In this section, I consider the elements of the FEMTI taxonomy of system characteristics and their portability to a similar taxonomy in FEIRI. It is, however, beyond the scope of this paper to completely flesh out those taxa that would necessitate modification when ported to FEIRI.

3.2.1 FEMTI System Characteristics Taxa that Port to FEIRI

There are many taxa in the FEMTI "System characteristics" taxonomy that port directly to FEIRI. Most, though not all, of these are derived from ISO 9126 (ISO/IEC, 2001). A notable exception is FEMTI taxon 2.7 "Cost", and its children. ISO does not address these, but as noted in FEMTI, "cost may play a major role in disbarring a system from detailed evaluation. It is therefore included here as part of the quality model." This is equally true for evaluation of MNM systems. Other taxa that carry over more or less directly include those under FEMTI 2.2 "Reliability", 2.3 "Usability", and 2.6 "Portability". Those under FEMTI 2.5 "Maintainability", 2.4 "Efficiency", and 2.1 "Functionality" are more of a mixed bag.

Of the taxa under 2.5 "Maintainability", 2.5.1 "Analyzability", 2.5.3 "Stability", 2.5.4 "Testability", and 2.5.5 "Maintainability compliance" all port directly over to FEIRI. Of the taxa under 2.1 "Functionality", quite a broad category, the following port directly to FEIRI: 2.1.4 "Interoperability", 2.1.5 "Functionality compliance" (including compliance with standards, specifically and notably data exchange standards), and 2.1.6 "Security" (in particular here, since the object of the processing is peoples names, data security is crucial to satisfy privacy laws).

3.2.1 FEMTI System Characteristics Taxa that Require Modification

Many of the taxa under FEMTI 2.1 "Functionality", specifically those under taxa 2.1.1 "Accuracy", 2.1.2 "Suitability", and 2.1.3 "Well-formedness" need to be substantially reworked to apply to MNM evaluation in their port to FEIRI. This is understandable, since much of the work of FEMTI was focused on fleshing out and specializing this section of the taxonomy specifically for MT evaluation. Given the cross-cultural nature of FEIRI, there are likely analogs to some of the taxa in this section, but some will be eliminated entirely in the port to FEIRI.

Likewise the children under taxon 2.4 "Efficiency", 2.4.1 "Time Behavior" that have to do with post-editing will most likely disappear from FEIRI, as will the as yet undefined 2.4.1.4.3 "Update Time". However, 2.4.1.2,

[6] An example of this would be the Slavic name "Pavel" appearing as "Bafil" by virtue of having passed through Modern Standard Arabic orthography, which does not have a letter equivalent to English "P" or "V".

"Pre-processing time" is still relevant to IR evaluation. In MNM evaluation specifically, it may have to be broken down even further, to include data cleanup, data loading, indexing, and any other specialized processing that must take place before the name data can be searched. "Codeset conversion" here listed as a post-processing step will likely be included as a consideration for preprocessing in FEIRI. Items under 2.4.2 "Resource Utilization" and 2.5.2 "Changeability" are also remarkably a propos for MNM evaluation, and would just have to be modified slightly to refer to MNM concepts rather than MT concepts (e.g. the idea of "lexicons" or "dictionaries" can be applied equally to data resources used by MNM systems, and some MNM systems have modifiable rules and/or parameters).

4. A Brief Note on Human Versus Automated Evaluation

In the call for papers, the organizers ask:

- Should we work on hybrid methodologies of automatic and human evaluation for certain technologies and not for others?

- Can we already envisage the integration of these approaches?

In answer to this, we might think of the canonical example of automatic evaluation – BLEU (Papineni *et al*, 2001). Even with this *automated* evaluation method it is necessary to put forth the human effort up front to develop one or (preferably) more reference translations to be used as the gold standard. Likewise, in MNM evaluation, if we imagine a methodology roughly like that described for TREC (Voorhees, 2001; Voorhees and Harman, 2000) or CLEF [7] (Peters, 2001; Peters and Braschler, 2001) consisting of human adjudication of pooled retrieval results from multiple systems, once we have those human relevance judgments, we have the equivalent of our gold standard translations, and have a method that is just as automatic for MNM evaluation as BLEU is for MT evaluation.

But, let's consider taking the automation one step further, into the space of the original human adjudications of the name matches. Imagine, for example, the use of an aggregate of MNM engines as an automatic method to bootstrap creation of ground truth for MNM evaluation. We already make the "closed world" assumption in the evaluation of name matching engines that we do – that is, if a particular query-result pair is not specifically listed as a true match in our ground truth set, it is assumed to be a non-matching pair. Assume now that we have a set of name matching engines. Assume further that all produce some score indicating the goodness of match between the query name and the name being returned, and that this score has an upper bound, which indicates an exact match between the query and the record returned. Then, in the extreme, we could at least add these exact match pairs to our ground truth as true matches. Presumably most, if not all, systems would have returned these matches, and even if this were not the case – perhaps due to flaws in some of the retrieval systems – at least one system should have returned each of the matches with a perfect score. So, the question, then, is how high must the similarity score for a given pair be, and how many engines need to have returned that pair in order for us to consider automatically adding it to our ground truth as a true name match? If we can determine the answer to this question, we will have gone some way in developng a semi- automatic way of creating at least partial ground truth for MNM evaluation. Taken further, this idea – in combination with research on the amount of ground truth necessary to produce useful results with stable system rankings – might greatly reduce the amount of effort needed to develop ground truth data to support evaluation of name matching engines.

[7] beginning in 1997 as a track in TREC, coordinated in Europe with the support of the European Commission since 2000, and now continuing under the auspices of the "TrebleCLEF Coordination Action" (http://www.trebleclef.eu/about.php).

5. Conclusion

In this paper, I have laid out considerations for the expansion of the Framework for the Evaluation of Machine Translation in ISLE (FEMTI) to a Framework for the Evaluation of Information Retrieval in ISLE (FEIRI), focusing on a specialized IR technology – the matching of multicultural personal names. I have concluded that, as the EAGLES 7-Step Recipe of task-focused evaluation has proven to be a useful and well-motivated evaluation method, not only for MT evaluation but also for multilingual name matching, and since that methodology that puts us squarely on the starting block of the first part of the FEMTI taxonomy, it does make sense to try to expand FEMTI to account for MNM evaluation.

I have noted several places where FEMTI must be updated in order to be applicable to the evaluation of multicultural name matching. I have also raised what I believe to be one crucial way in which MT evaluation differs from IR evaluation. That is, the quality characteristics for the output of IR seem much more clear cut than do those for the output of MT, as, therefore, do the mappings from those quality characteristics to methods for their measurement. In fact, this is likely a generalization that will hold in many cases as extension of FEMTI to evaluation of other HLT domains is considered. Because of this, the major strength of FEIRI may lie not in the mapping of evaluation requirements to system characteristics to metrics, as it does for FEMTI, but rather in the mapping from a combination of the characteristics related to sources of error and task definition (accompanied by the risk associated with failing to accomplish a certain level of accuracy on that task) to characteristics of the data to be used in the evaluation, and a threshold for accuracy to be attained as measured by the chosen metric.

References

.

Arehart, M., Miller, K.J., (2008). A Ground Truth Dataset for Matching Culturally Diverse Romanized Person

Names. Language Resources and Evaluation Conference 2008. Marrakesh, Morocco, 28 - 30 May 2008.

Braschler M., Di Nunzio G., Ferro N., Gonzalo J., Peters C., Sanderson M. (2007). From CLEF to TrebleCLEF: promoting Technology Transfer for Multilingual Information Retrieval. In Thanos C., Borri F., Launaro A. editors, *Working Notes of the Second DELOS Conference on Digital Libraries*, Pisa, Italy, 5-7 December 2007.

Hovy, E., King, M., Popescu-Belis, A. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 17:1, pp. 43-75.

ISO/IEC. 2001. International Standard ISO/IEC 9126-1. Software engineering -- Product quality -- Part 1: Quality model. International Organization for Standardization / International Electrotechnical Commission. Geneva.

Lloyd, D., Miller, K.J. (2007). An Optimization Approach for Identity Matching in the Federal Sector, *Joint Statistical Meetings 2007*, Salt Lake City, Utah, 29 July – 2 August 2007.

Miller, K.J., Arehart, M., Ball, C., Polk, J., Rubenstein, A., Samuel, K., Schroeder, E., Vecchi, E., Wolf, C. (2008). An Infrastructure, Tools and Methodology for Evaluation of Multicultural Name Matching Systems. In Language Resources and Evaluation Conference 2008. Marrakesh, Morocco, 28 - 30 May 2008.

Papineni, K.A., S. Roukos, T. Ward, W.J. Zhu. (2001). *BLEU: a method for automatic evaluation of machine translation*. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

Peters, C. (ed.) (2001). Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal 2000. *Lecture Notes in Computer Science* 2069, Springer 2001, 387p.

Peters, C., Braschler, M (2001). Cross-Language System Evaluation: the CLEF Campaigns. *Journal of the American Society for Information Science and Technology*, 52(12):1067-1072, 2001.

Voorhees, E.M. (2001). The Philosophy of Information Retrieval Evaluation. *Lecture Notes in Computer Science* 2406, pp. 355-370. London, UK: Springer-Verlag.

Voorhees, E.M. and D. Harman (2000). Overview of the Eighth Text REtrieval Conference (TREC-8). In D. Harman, ed., *The Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg, MD, USA.

Figure 3: FEMTI as Presented on the Web

Evaluating a Natural Language Processing Approach in Arabic Information Retrieval

Nasredine Semmar (1), Laib Meriama (1), Christian Fluhr (2)
(1) CEA, LIST, Laboratoire d'ingénierie de la connaissance multimédia multilingue
18 route du Panorama, BP6, FONTENAY AUX ROSES, F- 92265 France
E-mail: nasredine.semmar@cea.fr, meriama.laib@cea.fr

(2) NewPhenix
33 rue Galilée, PARIS, F- 75116 France
E-mail: christian.fluhr@new-phenix.com

Abstract

The purpose of information retrieval is to find all the relevant documents for a user's query in a collection of documents. Natural language processing for information retrieval consists in extracting concepts from the documents to be indexed and from the user's query. These concepts are used in matching and retrieval tasks. In this paper, we present the Arabic linguistic analyzer used in our cross-language search engine. As Arabic is a derivation based language in which morphology plays a significant role, we will focus particularly on the stemming process and the Part-Of-Speech tagging and their impact on the information retrieval effectiveness.

1. Introduction

Arabic is a derivation based language in which morphology plays a significant role (Zouari, 1989; Attia, 1999). Definite articles, conjunctions, particles and other prefixes can attach to the beginning of a word, and large numbers of suffixes can attach to the end. Moreover, in Arabic newspapers, texts are often completely or partially unvowelled and an unvowelled word can correspond to a set of potentially vowelled words having different meanings. For information retrieval, this abundance of forms, lexical variability, and orthographic alternatives, all result in a greater likelihood of mismatch between the form of a word in a query and the forms found in documents relevant to the query.

We present in section 2, the main components of our cross-language information retrieval system. In section 3, the linguistic analyzer, in particular, the morphological analyzer and the clitic stemmer are described. We present in section 4 the metrics used to evaluate the Arabic information retrieval system. We discuss in section 5 the results obtained after submitting in natural language short and long queries on a collection of documents related to water, sustainable development and tourism. Section 6 concludes our study and presents our future work.

2. Information retrieval

The purpose of information retrieval is to find all the relevant documents for a user's query in a collection of documents (Salton, 1983) and cross-language information retrieval aims to find relevant documents that are in a different language from that of the user's query (Grefenstette, 1998). Our cross-language information retrieval system (Semmar et al., 2005) is based on a weighted boolean model and is composed of the following modules:

- A linguistic analyzer which includes a morphological analyzer, a Part-Of-Speech tagger and a syntactic analyzer. The linguistic analyzer processes both documents to be indexed and queries to produce a set of normalized lemmas, a set of named entities and a set of nominal compounds with their grammatical tags.
- A statistical analyzer that computes for documents to be indexed concept weights based on concept database frequencies.
- A comparator which computes intersections between queries and documents and provides a relevance weight for each intersection.
- A reformulator to expand queries during the search. The expansion is used to infer from the original query words other words expressing the same concepts. The expansion can be in the same language (synonyms, hyponyms, etc.) or in different language.
- An indexer to build the inverted files of the documents on the basis of their linguistic analysis and to store indexed documents in a database.
- A search engine which retrieves the ranked and relevant documents from the indexes according to the corresponding reformulated query and then merges the results obtained for each language taking into account the original words of the query (before reformulation) and their weights in order to score the documents.

3. Linguistic analysis

Natural language processing for information retrieval consists in extracting concepts from the documents to be indexed and from the user's query. Our linguistic analyzer (Grefenstette et al., 2005) produces a set of normalized lemmas, a set of named entities and a set of nominal compounds. It is composed of a set of processing modules with their linguistic resources.

3.1 Linguistic processing modules

The linguistic analyzer is built using a traditional architecture involving separate processing modules:

- A Tokenizer which separates the input stream into a graph of words. This separation is achieved by an automaton developed for each language and a set of segmentation rules.
- A Morphological analyzer which looks up each word in a general full form dictionary. If these words are found, they are associated with their lemmas and all their grammatical tags. For Arabic agglutinated words which are not in the full form dictionary, a clitic stemmer (Larkey et al., 2002; Aljlayl & Freider, 2002) was added to the morphological analyzer. The role of this stemmer is to split agglutinated words into proclitics, simple forms and enclitics. The clitic stemmer proceeds as follows:
  1. Several vowel form normalizations are performed: the vowel symbols ً ٌ ٍ َ ُ ِ ّ ْ are removed, the characters آ إ أ are replaced by the character ا and the final characters ي ئ ؤ or ة are replaced by the characters ى یء وء or ه.
  2. All clitic possibilities are computed by using proclitics and enclitics dictionaries.
  3. A radical, obtained by removing these clitics, is checked against the full form lexicon. If it does not exist in the full form lexicon, re-write rules (Darwish, 2002) are applied, and the altered form is checked against the full form dictionary. For example, consider the token "بكرته" (with its ball) and the included clitics ب (with) and ه (its), the computed radical كرت does not exist in the full form lexicon but after applying one of the re-write rules, the modified radical "كرة" (ball) is found in the dictionary and the input token is segmented into root and clitics as: بكرته = ب + كرة + ه (with + its + ball).
  4. The compatibility of the grammatical tags of the three components (proclitic, radical, enclitic) is then checked. Only valid segmentations are kept and added into the graph of words.
- An Idiomatic Expressions recognizer which detects idiomatic expressions and considers them as single words for the rest of the processing. Idiomatic expressions are phrases or compound nouns that are listed in a specific dictionary. The detection of idiomatic expressions is performed by applying a set of rules that are triggered on specific words and tested on left and right contexts of the trigger. These rules can recognize contiguous expressions as "البَيْت الأبْيَض" (the white house). Non-contiguous expressions such as phrasal verbs are recognized too.
- A module to process unknown words by assigning to these words default linguistic properties based on features identified during tokenization (e.g. presence of Arabic or Latin characters, numbers, etc.).
- A Part-Of-Speech (POS) tagger which searches valid paths through all the possible tags paths using attested trigrams and bigrams sequences. The trigram and bigram sequences are generated from a manually annotated training corpus. They are extracted from a hand-tagged corpora of 13 200 Arabic words. If no continuous trigram full path is found, the POS tagger tries to use bigrams at the points where the trigrams were not found in the sequence. If no bigrams allow completing the path, the word is left undisambiguated. The accuracy of the Arabic Part-Of-Speech tagger is around 91%.

The following example shows the result of the linguistic analysis after Part-Of-Speech tagging of the Arabic sentence "خط أنابيب بين إيران وقطر لنقل المياه" (pipeline between Iran and Qatar to transport water).

```
(1) خط | خَـــطّ#L_NC_GEN
(2) أنـابـيـب | أُنْـبُـــــوب#L_NC_GEN
(3) بـيـن | بَـيْـــــنَ#L_PREP_AVEC_NOM
(4) إيـــران | إيـــرَان#L_NP
(5) و | وَ#L_CONJ_COORD
(6) قـطـر | قـطَـــر#L_NP_GEN
(7) لِ | ل#L_PREP_AVEC_NOM
(8) نـقـل | نَـقْـــل#L_NC_GEN
نُـقْـلَـــــة#L_NC_GEN
(9) ال | ال#L_DET_ARTICLE_DEF
(10) مـيـاه | مَـاء#L_NC_GEN
```

In this example, the Part-Of-Speech tagger has affected to the word "نقل" two lemmas "نَقْل" (transportation) and "نُقْلَة" (truckload) with the same grammatical tag "L_NC_GEN" which corresponds to a "Noun".

- A Syntactic analyzer which is used to split graph of words into nominal and verbal chain and recognize dependency relations (especially those within compounds) by using a set of syntactic rules. We developed a set of dependency relations to link nouns to other nouns, a noun with a proper noun, a proper noun with the post nominal adjective and a noun with a post nominal adjective. These relations are restricted to the same nominal chain and are used to compute compound words. For example, in the nominal chain "نقل المياه" (water transportation), the syntactic analyzer considers this nominal chain as a compound word "نقل_مياه" composed of the words "نقل" (transportation) and "مياه" (water).
- A Named Entity recognizer which uses name triggers (e.g., President, lake, corporation, etc.) to identify named entities (Abuleil & Evens, 2004). For example, the expression "الأوّل مِن شَهْر مَارس" (The first of March) is recognized as a date and the expression "قطر" (Qatar) is recognized as a location.
- A module to eliminate empty words which consists in identifying words that should not be used as search criteria and removing them. These empty words are identified using only their Part-Of-Speech tags (such as prepositions, articles, punctuations and some adverbs). For example, the preposition "ل" (for) in the agglutinated word "لنقل" (for transportation) is considered as an empty word.

• A module to normalize words by their lemmas. In the case the word has several lemmas, only one of these lemmas is taken as normalization. Each normalized word is associated with its morpho-syntactic tag. For example, normalization of the word "أنابيب" (pipelines) which is the plural of the word "أُنْبُوب" (pipeline) is represented by the couple (أُنْبُوب, Noun).

Table 1 illustrates the bag of words (results of linguistic analysis) of the sentence "خط أنابيب بين إيران وقطر لنقل المياه" after eliminating empty words and achieving normalization.

Simple words	Compound words	Named entities
خَطّ (line) أُنْبُوب (pipeline) نَقْل (transportation) مَاء (water)	خَطّ_أُنْبُوب نَقْل_مَاء	إيرَان (Iran) قَطر (Qatar)

Table 1: Results of the linguistic analysis of the sentence "خط أنابيب بين إيران وقطر لنقل المياه".

3.2 Linguistic resources

Linguistic resources involved in Arabic text processing are composed of:

• A set of rules for tokenizing words.

• A full form dictionary which contains 3 164 000 entries. Each entry is accented, normalized, associated to its unvowelled versions and has its possible Part-Of-Speech tags and linguistic features (gender, number, etc).

• Proclitics and enclitics dictionaries which have the same structure of the full form dictionary with vowelled and unvowelled versions of each clitic. They contain not only the individual proclitics and enclitics but all valid concatenations of proclitics as well. No linguistic properties are assigned to concatenations of clitics. Each component of concatenated particles has its own linguistic properties. There are 77 and 65 entries respectively in each dictionary.

• A set of Part-Of-Speech n-grams (bigrams and trigrams from hand-tagged corpora) that are used for Part-Of-Speech tagging.

• A set of rules for shallow parsing of sentences to extract compounds from the input text.

• A set of rules for the identification of named entities: gazetteers and contextual rules that use special triggers to identify named entities and their type.

4. Evaluation metrics

In evaluation campaigns such as TREC or CLEF, the relevance is formulated as a binary score according to how well does the document answer the user query. A document obtains the score 1 if it is considered as relevant or the score 0 if it is considered as not relevant. But for human assessment, it is difficult to decide if a document is wholly relevant or not because it can more or less deal with information contained in the query. We can measure the relevance according to different levels, but this approach is highly priced and can't be applied in a big scale campaign.

The evaluation of information retrieval effectiveness can use different metrics. Precision and recall are the most common used measures:

• Precision: corresponds to the proportion of relevant documents, with regard to the number of all documents returned by the information retrieval system. A precision equal to 1 is obtained in an exceptional case when the system has only one relevant document to propose and when this document is the only one corresponding to the user's query.

$$\text{Precision} = \frac{|A_r \cap A|}{|A|}$$

Where A_r corresponds to the set of the relevant documents and A corresponds to the set of the retrieved documents.

• Recall: corresponds to the proportion of relevant documents returned by the system, with regard to the number of all relevant documents in the database. A Recall equal to 1 is obtained when a system return all documents of the database as a response to the user's query.

$$\text{Recall} = \frac{|A_r \cap A|}{|A_r|}$$

• F-measure: establishes a compromise between precision and recall. The F-measure corresponds to the weighted harmonic mean of precision and recall. The formula to compute the F-measure is as follows:

$$\text{F - measure} = 2 * \frac{\text{Pr}\,ecision * \text{Re}\,call}{\text{Pr}\,ecison + \text{Re}\,call}$$

5. Experimental results and discussion

Our cross-language search engine has been tested on a multilingual corpora provided by partners of European project ALMA (Semmar & Fluhr, 2004). This corpus contains for each language (Arabic, English and French) 50 non-parallel documents related to sustainable development, water and eco-tourism. We used the track guidelines of TREC 2002 to evaluate the Arabic search engine. We built manually a set of 50 Arabic short queries, a set of 50 Arabic long queries and a list of these queries and their corresponding relevant Arabic documents (human judgments).

The following example shows a query according to TREC 2002 guidelines.

```
<top>
<num>50</num>
<title>إدارة موارد المياه</title>
<desc> تعزيز قدرات إدارة موارد المياه وإشراك القطاع الخاص فى أنشطة إعادة الاستخدام</desc>
<narr></narr>
</top>
</topics>
```

We launched two runs: one with short queries (content of the tag "<title>") and the other with long queries (content of the tag "<desc>"). In the two cases, we took into account the 50 documents returned by the search engine and we used the *trec_eval* application to evaluate the results.

Tables 2 and 3 show respectively recall and precision measures for short and long queries.

Recall	Precision	F-measure
0,1	0,76	0,17
0,2	0,74	0,31
0,3	0,72	0,42
0,4	0,70	0,51
0,5	0,69	0,58
0,6	0,64	0,62
0,7	0,61	0,65
0,8	0,60	0,68
0,9	0,57	0,70

Table 2: Recall and precision measures for short queries.

Recall	Precision	F-measure
0,1	0,93	0,18
0,2	0,91	0,32
0,3	0,90	0,45
0,4	0,86	0,54
0,5	0,85	0,62
0,6	0,79	0,68
0,7	0,75	0,72
0,8	0,75	0,77
0,9	0,70	0,79

Table 3: Recall and precision measures for long queries.

The results we obtained (Tables 2 and 3) show that for the same value of the recall the precision is better when the query is long. This is due to the fact that Part-Of-Speech tagging assigns correct grammatical tags to the words of the query when the query is long.

In order to evaluate how the linguistic processing can improve the results particularly when the request is long, we compared the returned relevant documents with the human judgments.
The task consisted in submitting queries with just one word. The results have been obtained with two Part-Of-Speech taggers. The first one chooses just one grammatical tag among all the other valid tags and the second keeps all the valid tags.

For example, for the query containing only the word "قطر" (Qatar, the country), the search engine returned relevant documents when we used the second Part-Of-Speech tagger and did not return these relevant documents when we used the first POS tagger.

To understand these results, we analyzed the output produced by the linguistic processing of the query. The word "قطر" which has no vowels in the input texts (documents and query), has several grammatical tags in the dictionary according to the vowels it can have: قَطَّرَ (purify) as a verb, "قُطْرُ" (Country) as a noun and "قَطَر" (Qatar) as a proper noun. When the first Part-Of-Speech tagger is used, the word "قطر" obtains the tag "Verb". The same word in the relevant document is considered as "Proper Noun". Moreover, as their two lemmas are different, the comparator did not find a similarity between the query and the document. When the second Part-Of-Speech tagger is used, all the tags of the word are kept with their lemmas and the match is then possible.

According to these observations, we can deduce that the use of a Part-Of-Speech tagger which proposes all the valid tags of the words and their lemmas can improve the results of information retrieval even if queries are short.

In addition, the comparator of the search engine uses only the lemmas of the words to retrieve the relevant documents and it does not use the grammatical tags of these words. For example, for the query containing the word "مارس" (March, the month), the search engine returned documents which are not relevant. After analyzing the results of the linguistic processing of the query and the documents we found that the first Part-Of-Speech tagger has assigned the grammatical tag "Verb" to the word "مارس" of the query and the grammatical tags "Verb" and "Noun" for the same word which is present in two documents. These tags are correct but the search engine returned the irrelevant document in which the word "مارس" has the tag "Verb". This is due to the fact that the compartor of the search engine uses only the lemmas of the words to compute intersections between queries and documents.

On the other hand, we noticed that the search engine did not return documents containing the word "آذار" which is a synonym of the word "مارس". Indeed, the current version of our search engine does not use a dictionary of synonyms to reformulate Arabic words.

Another module of the linguistic analysis which improves information retrieval effectiveness is the clitic stemmer.

For example, for the words "المياه", "الماء", "مياه" or "ماء" (different surface forms of the word "water"), the clitic stemmer computes all the clitic possibilities and proposes the stem "ماء" as a lemma for these different surface forms.

Furthermore, we compared our results with some of the current commercial information retrieval tools (Alltheweb and Google) by querying these tools with the same words "المياه", "الماء", "مياه" or "ماء". We noted that documents returned by these search engines depend on the used surface form. This is due to the fact that these tools do not use stemming to extract the stems from the Arabic agglutinated words of the query and the indexed documents (Abdelali et al., 2004).

6. Conclusion and future work

Our experiments showed that stemming Arabic words of queries and documents sentences contributes significantly to improve information retrieval. These experiments confirmed results obtained by Larky and al. (2002) with their light stemmer. On the other hand, our experiments demonstrated the importance of Part-Of-Speech tagging particularly on short queries. In order to confirm these results, we will evaluate in a future work our search engine on the Arabic TREC 2002 corpus.

7. References

Abdelali, A., Cowie, J., Soliman, H. S. (2004). Arabic Information Retrieval Perspectives. In *Proceedings of the Conference on Natural Language Processing*. Fez.

Abuleil, S., Evens, M. (2004). Named Entity Recognition and Classification for Text in Arabic. In *Proceedings of the 13th International Conference on Intelligent & Adaptive Systems and Software Engineering.* Nice, pp. 89--94.

Aljlayl, M., Frieder, O. (2002). On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach. In *Proceedings of the eleventh international conference on Information and knowledge management.* McLean, VA, pp. 340--347.

Attia, M. (1999). A large-Scale Computational Processor of Arabic Morphology and Applications. M.S. Thesis in Computer Engineering. Cairo University.

Darwish, K. (2002). Building a Shallow Arabic Morphological Analyzer in One Day. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* Philadelphia, PA, pp. 47--57.

Grefenstette, G. (1998). Cross-language Information Retrieval. Kluwer Academic Publishers. Amherst, MA.

Grefenstette, G., Semmar, N., Elkateb-Gara, F. (2005). Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information Processing and Information Retrieval Applications. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics.* Ann Arbor, MI, pp. 31--38.

Larkey, L. S., Ballesteros, L., Connell, M. E. (2002). Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.* Tampere, pp. 275--282.

Salton, G., Michael, J. M. (1983). Introduction to modern information retrieval. McGraw Hill. New York.

Semmar, N., Fluhr, F. (2004). Multilingual Search Engine implementation. Final report of ALMA project, EURO-MED programme. DG XIII, Commission of the European Union.

Semmar, N., Elkateb-Gara, F., Laib, M., Fluhr, F. (2005). A Cross-language information retrieval system based on linguistic and statistical approaches. In *Proceedings of the Deuxième Congrès International sur l'Ingénierie de l'Arabe et l'Ingénierie de la Langue.* Alger, pp. 114--125.

Zouari, L. (1989). Construction Automatique d'un Dictionnaire Orienté vers l'Analyse Morpho-syntaxique de l'Arabe Ecrit Voyellé ou Non Voyellé. Thèse de Doctorat en Informatique. Université Paris-sud, Centre d'Orsay.

A Review of the Benefits and Issues of Speaker Verification Evaluation Campaigns

Asmaa El Hannani[1], **Jean Hennebert**[2,3]

[1]Deptartment of Computer Science, University of Sheffield, UK
[2]HES-SO, Business Information Systems, TechnoArk 3, CH-3960 Sierre, Switzerland
[3]University of Fribourg, Bd de Prolles 90, CH-1700 Fribourg, Switzerland
asmaa.elhannani@sheffield.ac.uk, jean.hennebert@hevs.ch

Abstract

Evaluating speaker verification algorithms on relevant speech corpora is a key issue for measuring the progress and discovering the remaining difficulties of speaker verification systems. A common evaluation framework is also a key point when comparing systems produced by different labs. The speech group of the National Institute of Standards and Technology (NIST) has been organizing evaluations of text-independent telephony speaker verification technologies since 1996, with an increasing success and number of participants over the years. These NIST evaluations have been recognized by the speaker verification scientific community as a key factor for the improvement of the algorithms over the last decade. However, these evaluations measure exclusively the effectiveness in term of performance of the systems, assuming some conditions of use that are sometimes far away from any real-life commercial context for telephony applications. Other important aspects of speaker verification systems are also ignored by such evaluations, such as the efficiency, the usability and the robustness of the systems against impostor attacks. In this paper we present a review of the current NIST speaker verification evaluation methods, trying to put objectively into evidence their current benefits and limitations. We also propose some concrete solutions for going beyond these limitations.

1. Introduction

Speaker verification consists in verifying a person's claimed identity. It is a subfield of speaker recognition that comprises all of the many different tasks of distinguishing people on the basis of their voices.

Speaker verification is also a subfield of biometric technologies. Biometrics, which bases the person authentication on the intrinsic aspects of a human being, appears as a viable alternative to more traditional approaches such as keys, badges, magnetic cards or memorized passwords. Biometric person authentication could be done using various modalities such as fingerprints, face, speech, dynamic signature, iris, hand geometry or keystroke dynamics. As a biometric modality, speech has a number of advantages and potentialities in comparison to the other modalities. Speech does not require any physical contact with the acquisition device and so is considered lowly intrusive by users. Moreover, in some cases (over the telephone, the radio, in the dark ...), speech is often the only available modality to recognize the identity of a person.

There are two main tasks of speaker recognition; speaker identification and speaker verification. The difference between these two tasks rests mainly on the type of decision that should be made. Usually, they are both based on the same modelling technologies (Wan and Campbell, 2000; Reynolds, 1995). The **speaker identification** task consists in determining, from a sequence of speech samples, the identity of an unknown person among N recorded speakers, called reference speakers. The identification answers the question "Whose voice is this?". This process gives place to N possible results. The **speaker verification** (also referred as speaker detection) aims to determine if a person, who claims to be a target speaker[1], is or is not this speaker. The decision will be either an acceptance or a rejection. The verification answers the question, "Am I who I claim to be?". If the person is not a target speaker, he is called an impostor.

Evaluating speaker recognition algorithms on relevant speech corpora is a key issue for measuring the progress and assessing the difficulties of speaker verification systems. In an ongoing effort to support research in text-independent speaker recognition technologies, NIST has been conducting annual speaker recognition evaluations. The aim of these evaluations is to provide common framework (data, rules and scoring) to allow focused technology development and meaningful comparison of techniques and approaches. However, these evaluations measure exclusively the effectiveness in term of performance of the systems, assuming some conditions of use that are sometimes far away from any real-life commercial context for telephony applications. Other important aspects of speaker verification systems are also ignored by such evaluations, such as the efficiency, the usability and the robustness of the systems against impostor attacks. Finally, using a similar evaluation framework for consecutive years have made converge many labs into using similar kind of algorithms (mostly GMM based) where system differences are essentially linked to the ability of the labs to aggregate large quantity of data used for training normalization and background components of the system.

2. Speaker Verification Evaluation

2.1. Performance Factors

Speaker verification performance is dependent upon many different factors that could be grouped in the following categories:

[1]Also referred in the literature as true, reference or client speaker

- **Intra-speaker Variabilities**: Usually the speaker model is obtained using a limited amount of speech data that characterizes the speaker at a given time and situation. However, the voice can change in time due to aging, illness, emotions, tiredness and potentially other factors. For these reasons, the speaker model may not be representative of the speaker in all his/her potential states. Variabilities may not all be covered, which affect negatively the performance of the speaker verification systems. To deal with this problem, incremental enrollment techniques can be used in order to include the short and long-term evolution of the voice (see for example (Barras et al., 2004)).

- **Mismatch Factors**: The mismatch in recording conditions between the training and testing is the main challenge for automatic speaker recognition, specially when the speech signal is acquired on telephone lines. Differences in the background noise, in the telephone handset, in the transmission channel and in the recording devices can, indeed, introduce variabilities over the recording and decrease the accuracy of the system. This is mainly due to the statistical models that do not capture only the speaker characteristics but also the environmental ones. Hence, the system decision may be biased if the verification environment is different from the enrollment. The features and score normalization techniques (e.g. (Pelecanos and Sridharan, 2001; Reynolds et al., 2003; Auckenthaler et al., 2000; Reynolds et al., 2000)) are useful to make speaker modelling more robust to recording conditions. The high-level features (e.g. (Reynolds et al., 2003; Campbell et al., 2003; El Hannani and Petrovska-Delacrétaz, 2007)) are also important because they are supposed to be more robust to mismatched conditions.

- **Amount of Speech Data**: The amount of training data available to build the speakers model and to test it has also a large impact on the accuracy of the systems. This was confirmed during the NIST Speaker Recognition Evaluation (SRE) evaluations (Martin and Przybocki, 2004), where it has been shown that the duration and number of sessions of enrollment and verification affect the performance of the speaker verification systems.

2.2. Performance Measures

The performance of any speaker recognition system is evaluated in function of the error rate. There are two types of errors that occur in a verification task: the false acceptance when the system accepts an impostor and the false rejection when the system rejects a valid speaker. Both types of errors depend on the decision threshold. With a high threshold, the system will be highly secured. In other words, the system will make very few false acceptances but a lot of false rejections. If the threshold is fixed to a low value, the system will be more convenient to the users making few false rejections and lots of false acceptances. The rates of false acceptance, R_{FA}, and false rejection, R_{FR}, are then functions of the threshold and define the operating point of the system. They are calculated as follows:

$$R_{FA} = \frac{number\ of\ false\ acceptances}{number\ of\ impostors\ access} \quad (1)$$

$$R_{FR} = \frac{number\ of\ false\ rejections}{number\ of\ targets\ access} \quad (2)$$

These rates are normally estimated on the development set and are further used to compute the Detection Cost Function (DCF). This cost function is a weighted measure of both false acceptance and false rejection rates:

$$DCF = C_{FR}P_{tar}R_{FR} + C_{FA}P_{imp}R_{FA} \quad (3)$$

where C_{FR} is the cost of false rejection, C_{FA} is the cost of false acceptance, P_{tar} is the a priori probability of targets and P_{imp} is the a priori probability of impostors.
The DCF is the most used measure to evaluate the performances of operational speaker verification systems. The smaller is the value of the DCF, the better is the system for the given application and conditions. Thus, the decision threshold is usually optimized in order to minimize the DCF. This optimization is often done during the development of the system on a limited set of data.
Another popular measure is the Equal Error Rates (EER). It represents the error at the threshold which gives equal false acceptances and false rejections rates. The EER is not interpretable in function of the cost but still widely used as a reference indication of the performance of the system.

2.3. Detection Error Tradeoff Curve

The measures presented before evaluate the performances of the system in a single operating point. However, representing the performance of the speaker verification system over the whole range of operating points is also useful and can be achieved by using a performance curve. The Detection Error Tradeoff (DET) curve (Martin et al., 1997), a variant of the Receiver Operating Characteristic (ROC) curve (Egan, 1975), has been widely used for this purpose. In the DET curve the R_{FA} is plotted as a function of the R_{FR} and the axis follow a normal deviate scale. The points of the DET curve are obtained by varying the threshold T. This representation allows an easy comparison of the performances of the systems at different operating points. The EER appears directly on this curve as the intersection of the DET curve with the first bisectrix.

2.4. Speech Corpora and benchmarks

There has been a plethora of speaker verification algorithms and technologies proposed by the scientific communities and commercial vendors. Evaluating speaker recognition algorithms on relevant speech corpora has become a key factor for measuring the progress and detecting difficulties of speaker recognition systems. A survey of standard speech corpora that are suitable for the development and evaluation of speaker recognition systems can be found in (Godfrey et al., 1994; Campbell and Reynolds, 1999). The main suppliers of these corpora are the European Language Resources Association (ELRA)[2], the Linguistic Data

[2] http://www.elra.info

Consortium (LDC)[3], and the Oregon Graduate Institute (OGI)[4]. The most used corpora for speaker recognition are listed in Table 1.

Corpora	Supplier
SIVA PolyVar POLYCOST	ELRA
Switchboard I & II & Cellular TIMIT & NTIMIT & HTIMIT & CTIMIT NIST SREs Subsets Fisher KING YOHO SPIDRE CSLU TSID	LDC
Speaker Recognition Corpus	OGI

Table 1: Speaker recognition corpora and their suppliers.

Methodologies to benchmark the many different speaker recognition approaches have soon been developed on top of the available corpora. Generally speaking, one can classify benchmark methodologies as explained in (Cappelli et al., 2006) and as illustrated on Figure 2.4.:

- **In-house evaluation with self-defined test**: The testing protocol is self-defined on a privately owned database. Often, the recording conditions are controlled by the lab. As a consequence, results are not easily reproducible by a third party. The door is also open to data manipulation such as selection of speakers, discarding of outliers, etc. From an algorithmic point of view, problems of over-fitting the speaker data may also arise, i.e. the algorithms become too specific to a given data set. This is especially true if the evaluation protocol is not organized into independent development and evaluation sets.

- **In-house evaluation with existing benchmark**: The testing protocol and the corpora are publicly available. Assuming that the pre-defined evaluation protocols are strictly followed, results of algorithms executed on the data are comparable across sites and publications. Some existing corpora provide a defined evaluation procedure such as TIMIT (Reynolds et al., 1995), POLYCOST (Melin and Lindberg, 1996)(Hennebert et al., 2000), KING (Reynolds, 1994) or YOHO (Campbell, 1995). However, the risk of overfitting is definitively remaining as the protocols are often not organized into independent development and evaluation sets.

- **Independent weakly supervised evaluation**: The testing protocol is defined by an independent institute and the data, supposedly unseen by the participants, are made available just before the beginning of the test. Data samples are unlabeled (no ground truth) and the participant provide the evaluator with the results of the algorithms within given time constraints. All NIST Speaker Recognition Evaluations[5] fall in this category. For NIST, the unseen data are provided to the participants in a pretty large quantity to minimize the risk of participants willing to listen to the waveforms and manually perform the verification. However, the quantity of data is so large the time constraints are so strict that participating to such evaluations requires resources (human and cpu) that are often not available in participating labs.

- **Independent supervised evaluation**: The data are here completely sequestered by the evaluator. The participant provides the evaluator with a full solution to run the tests, including hardware and software. The evaluator can then better control the evaluation and the risk of human intervention is minimized. The drawback of the approach is that the evaluation can only be performed in terms of accuracy and not in terms of cpu or memory footprint.

- **Independent strongly supervised evaluation**: The participant provides here the evaluator with a software only solution that is run on the evaluator hardware. Recently, the 2007 Biometric Multimodal Evaluation Campaign was organized following the independent strongly supervised evaluation scheme described above. Speaker verification was present as part of the talking face evaluation (Fauve et al., 2008). Keeping the same hardware allows performing full comparisons in terms of performance, cpu and memory footprint. However, the drawback lies in the extra difficulty for the participant to modify its software so that it complies with a given input-output framework. The costs in terms of time and resources are also much larger on the side of the evaluator.

3. Overview of NIST Speaker Recognition Evaluation

3.1. History

The speech group of the National Institute of Standards and Technology (NIST) has now been organizing evaluations of speaker recognition technologies since 1996 with an increasing success over the years. The NIST Speaker Recognition Evaluation (SRE) campaigns varied from 1996 to 2006 in term of tasks and corpora used (Martin and Przybocki, 2004; Przybocki et al., 2006). The speaker detection (verification) task has remained the primary task over the years. However the evaluations have started including some other tasks such as speaker tracking and speaker segmentation. NIST included speaker tracking task between 1999 and 2001 and the speaker segmentation task between 2000 and 2002. The dadasets used for the evaluation have also changed to include different handsets, transmission types and languages. Table 2 summarizes the evolution of NIST SRE regarding the corpus, tasks, and training/testing durations.

[3] http://www.ldc.upenn.edu

[4] http://cslu.cse.ogi.edu

[5] http://www.nist.gov/speech/tests/spk

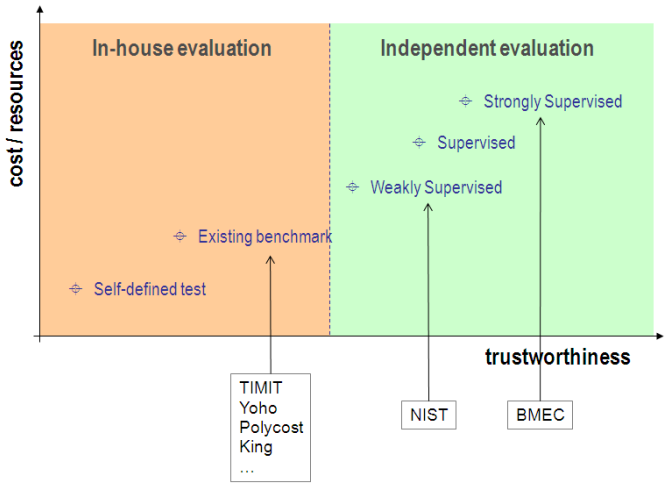

Figure 1: Classification of benchmark evaluations (after (Cappelli et al., 2006).

3.2. Evaluation Methodology

In NIST evaluations, common data-sets, standard measurements of error and evaluation protocol are provided to each participating laboratory. Each evaluation is followed by a workshop so that researchers can compare their submitted results and highlight problems that require further research. For each evaluation, NIST specifies the evaluation tasks and rules in its evaluation plan. The evaluation plan defines the datasets that participants may use during the evaluation procedure. This includes the training data used to train the system, the development data on which participants test and tune their systems and finally the evaluation data to perform the final tests of the system that will be scored by NIST. The evaluation plan includes also the dates for NIST to release the different types of data to the participants, and for the participants to submit their results to NIST.

The submitted results are scored by NIST and the final performances are made available to the participants few weeks after the submission. NIST uses the cost function described above as the basic performance measure (see equation 3). The cost of false rejection C_{FR} has been set as 10 and the cost of false acceptance C_{FA} to 1. The a priori probability of a target P_{tar} has been assigned the value 0.01 and P_{imp} the value 0.99.

3.3. Limitations

NIST SREs have been recognized by the speaker verification scientific community as a key factor for the improvement of the algorithms over the last decade. Nevertheless, they present some limitations:

First, NIST SREs are mostly relevant for applications where the interest is to find if a target speaker is present in a given test speech signal. The mode is fully text independent, i.e. there is no a priori control of what the speaker is saying. Such applications include surveillance applications as well as application for speaker segmentation, clustering or database annotations. Speaker verification has a large potential for commercial applications but in order to be convenient for the users, such systems need to be functional with short training and testing data and with controls to impeach replay attacks. Commercial applications are then mostly based on text-dependent systems and more specifically on text-prompted scenarios, where the speaker is requested to repeat a given utterance. Text-dependent or text-prompted scenarios have never been included in NIST SREs. This fact could explain the lack of commercial vendors participation to those evaluations.

Second, the quality of NIST data may skew the recognition performances of the systems. Indeed speakers could show variabilities due to factors such as topic of conversation, familiarity level with the interlocutor, etc. However the data used by NIST does not control such parameters. For example MIXER corpus was collected in order to support the US government needs with emphasis on forensic-style problem. The main goal was to improve the FBI's Forensic Automatic Speaker Recognition prototype which is designed to be text-independent, channel-independent and to recognize criminals and terrorist talking in different languages (Cieri et al., 2004). For this reason, the focus was more on the languages and channels conditions. This is certainly of interest to the program sponsors but not to the majority of researchers.

Third, the robustness of the systems against impostor attacks is not taken into account by NIST SRE. All impostors access used by NIST are done with zero-effort using so-called random impostures. This means that the impostors attacks are just simulated by testing the target voice against another speaker which is not realistic. Real impostors will of course put more efforts in order to attack the system. This could be by attempting to change their voices, playing a pre-recorded voice or using a text-to-speech system tuned to reproduced voice characteristics close to the one of the target speaker.

Finally, other important aspects of speaker verification systems are also ignored by NIST evaluations, such as the efficiency and the usability. Most of the systems presented in NIST SREs workshops are far away from any real-life commercial context for telephony applications. They require either lots of training data or lots of processing time which is ineffective from the usability point of view.

4. Discussions

There is actually an increasing interest in telephony based speaker recognition applications. Most of the existing commercial applications are text-dependent or text-prompted. So there is an urgent demand to collect relevant databases with which researchers can make a meaningful comparison of different state-of-the-art approaches and assess the progress they could make in this field. Also, and contrary to the NIST SRE data, the acquisition conditions should be as close as possible to real life conditions as encountered in commercial applications. This means short training and testing data, multichannel, mismatched recording conditions, text-prompting, incremental enrollment, etc.

More advanced speaker impostor technologies could also be used such as the impostor voice transformation (Perrot et al., 2005; Matrouf et al., 2006). This technique has been shown to increase the false acceptance rates with the advan-

tage to be low cost in terms of time and human efforts. In the same idea, the detection of replay attacks would also be an interesting challenge oriented towards improving the rejection of impostors. None of these directions have been currently taken by large scale evaluation benchmarks in speaker verification.

Usability tests are also important when considering real-life applications that imply interaction with the user. However, we believe that including such aspects in large scale evaluations are not tractable if it implies an analysis of a life user reaction in front of the system. This is especially true for evaluations such as NIST SRE where there is an increasing number of participants. Nevertheless, one could include more objective criteria oriented towards usability. For example, the duration of the test segments needed by the system to be effective could give an indication on the amount of effort required from the users and would therefore be linked to the acceptance of the system. Also, the system reactiveness could be measured looking at the real time factor of the algorithms. The reactiveness is indeed linked to a good acceptance of the system. Finally, the robustness of the systems in front of longer term intra-speaker variabilities should also be considered because this is the key factor of the user satisfaction.

5. Conclusions

In this paper we presented a review of the current ways to evaluate speaker verification systems, putting an emphasis on the NIST Speaker Recognition Evaluation methods. We tried to put objectively into evidence the current benefits and limitations of such evaluations. NIST SRE is the largest speaker recognition event in which the participating labs can make meaningful comparison of their different approaches with a common evaluation framework and pre-defined protocols. However, the tasks adopted by NIST SRE are, according to us, pretty far away from real-life application and are mostly relevant for applications such as surveillance or mining. Therefore, we believe there is a need of databases and evaluations that are closer to commercial applications of speaker verification systems (short training and testing data, multichannel, mismatched recording conditions, text-prompting, incremental enrollment, etc.). Also, evaluations organizers should include other criteria in order to develop a more user-centered approach. Finally, evaluations should attempt to evaluate stronger forgery scenarios than random impostures as the ability of the system to reject impostor attacks is also an important feature for many applications.

6. References

R. Auckenthaler, M.J. Carey, and H. Llyod-Thomas. 2000. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10.

C. Barras, S. Meignier, and J.-L. Gauvain. 2004. Unsupervised online adaptation for speaker verification over the telephone. *In the proceedings of the IEEE Workshop on Speaker and Language Recognition (Odyssey)*, June.

J.P. Campbell and D. Reynolds. 1999. Corpora for the evaluation of speaker recognition systems. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March.

J.P. Campbell, D. Reynolds, and R. Dunn. 2003. Fusing high- and low-level features for speaker recognition. *In the proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, September.

J.P. Campbell. 1995. Testing with the yoho cd-rom voice verification corpus. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 341–344, May.

R. Cappelli, D. Maio, D. Maltoni, J.L. Wayman, and A.K. Jain. 2006. Performance evaluation of fingerprint verification systems. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 28(1):3–18, January.

C. Cieri, J.P. Campbell, H. Nakasone, D. Miller, and K. Walker. 2004. The mixer corpus of multilingual, multichannel speaker recognition data. *In the proceedings of the International Conference on Language Resources and Evaluation*, May.

J. Egan. 1975. *Signal Detection Theory and ROC Analysis*. Academic Press.

A. El Hannani and D. Petrovska-Delacrétaz. 2007. Fusing acoustic, phonetic and data-driven systems for text-independent speaker verificationn. *In the proceedings of Interspeech*, August.

B. Fauve, H. Bredin, W. Karam, F. Verdet, A. Mayoue, G. Chollet, J. Hennebert, R. Lewis, J. Mason, C. Mokbel, and D. Petrovska. 2008. Some results from the biosecure talking face evaluation campaign. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

J. Godfrey, D. Graff, and A. Martin. 1994. Public databases for speaker recognition and verification. *ESCA Workshop on Automatic Speaker Recognition Identification and Verification*, pages 39–42, April.

J. Hennebert, H. Melin, D. Petrovska, and D. Genoud. 2000. Polycost: A telephone-speech database for speaker recognition. *Speech Communication*, 31(2-3):265–270, June.

A.F. Martin and M. Przybocki. 2004. Nist speaker recognition evaluation chronicles. *In the proceedings of the IEEE Workshop on Speaker and Language Recognition (Odyssey)*, June.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. 1997. The det curve in assessment of detection task performance. *In the proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 4:1895–1898, September.

D. Matrouf, J.-F. Bonastre, and C. Fredouille. 2006. Effect of voice transformation on impostor acceptance. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May.

H. Melin and J. Lindberg. 1996. Guidelines for experiments on the polycost database. *Proc. COST250 Workshop on The Application of Speaker Recognition Technologies in Telephony*, pages 59–69, November.

J. Pelecanos and S. Sridharan. 2001. Feature warping for

robust speaker verification. *In the proceedings of the IEEE Workshop on Speaker and Language Recognition (Odyssey)*, June.

P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet. 2005. Voice forgery using alisp: Indexation in a client memory. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1, March.

M. Przybocki, A.F. Martin, and A.N. Le. 2006. Nist speaker recognition evaluation chronicles - part 2. *In the proceedings of the IEEE Workshop on Speaker and Language Recognition (Odyssey)*, June.

D.A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. OLeary, and B. A. Carlson. 1995. The effects of telephone transmission degradations on speaker recognition performance. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May.

D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. 2000. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, January/April/July.

D. Reynolds, W. Andrews, J.P. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, J. Jones, and B. Xiang. 2003. The supersid project: Exploiting high-level information for high-accuracy speaker recognition. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April.

D.A. Reynolds. 1994. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(3):639–643.

D.A. Reynolds. 1995. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(1):91–108, August.

V. Wan and W. Campbell. 2000. Support vector machines for speaker verification and identification. *In proceedings of the IEEE Signal Processing Society Workshop*, 2:775–784.

Year	Corpus	Tasks	Training duration	Testing duration
1996	SWBD I	Speaker detection	2 minutes	3, 10 and 30 seconds
1997	SWBD II p1	Speaker detection	2 minutes	3, 10 and 30 seconds
1998	SWBD II p2	Speaker detection	2 minutes	3, 10 and 30 seconds
1999	SWBD II p3	Speaker detection Speaker tracking	1 minute	30 seconds
2000	SWBD p1 and p2 AHUMADA	Speaker detection Speaker tracking Speaker segmentation	2 minutes	15 to 45 seconds
2001	2000 dataset SWBD I	Speaker detection Speaker tracking Speaker segmentation	2 minutes	15 to 45 seconds
2002	SWBD cellular p1 SWBD p2 and p3 FBI Voice DB	Speaker detection Speaker segmentation	2 minutes, 1, 2, 4, 8 and 16 conversations	15 to 45 seconds and 1 conversation
2003	SWBD cellular p2	Speaker detection	2 minutes, 1, 2, 4, 8 and 16 conversations	15 to 45 seconds and 1 conversation
2004	MIXER	Speaker detection	10, 30 seconds, 5, 15, 40, and 80 minutes	10, 30 seconds and 5 minutes
2005	2004 dataset	Speaker detection	10 seconds, 5, 15, and 40 minutes	10 seconds and 5 minutes
2006	New MIXER data 2005 dataset	Speaker detection	10 seconds, 5, 15, and 40 minutes	10 seconds and 5 minutes

Table 2: History of the NIST Speaker Recognition Evaluations campaigns. The training and testing duration are reported for the speaker detection task only.

Field Testing of an Interactive Question-Answering Character

Ron Artstein, Sudeep Gandhe, Anton Leuski and David Traum

Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292, USA
artstein,gandhe,leuski,traum [at] ict.usc.edu

Abstract

We tested a life-size embodied question-answering character at a convention where he responded to questions from the audience. The character's responses were then rated for coherence. The ratings, combined with speech transcripts, speech recognition results and the character's responses, allowed us to identify where the character needs to improve, namely in speech recognition and providing off-topic responses.

Figure 1: SGT Star

Figure 2: SGT Star in a live demo

1. Background

We created Sergeant Star, a virtual question-answering character, for the U.S. Army Recruiting Command as a hi-tech attraction and an information source about the Army. He is a life-size character built for use in mobile exhibits, who accepts speech input and responds with pre-recorded voice answers (Figure 1). SGT Star is based on technology similar to that used in previous efforts (Leuski et al., 2006; Robinson et al., 2008), which treats question answering as an information retrieval problem: given a natural-language question, the character should retrieve the most appropriate answer from a list of available responses. A user's question is transformed to text through an Automatic Speech Recognition (ASR) module, and possible responses are ranked by a statistical classifier trained on a set of questions linked to responses. If the top-ranked response exceeds a preset threshold, the character utters that response; if no response reaches the threshold, the character makes a random choice among a set of predefined "off-topic" responses (like "Take the gum out of your mouth"). The system also includes provisions for detecting repetitive questions, and prompting the user if they are unable to come up with questions that the character can answer.

2. Setting

SGT Star was designed to be part of a mobile exhibit, so we evaluated his performance on the road, at the National Future Farmers of America Convention on 24–27 October 2007 in Indianapolis, where SGT Star was part of the U.S. Army exhibit (Figure 2 shows a similar installation at Fort Knox, Kentucky). Since SGT Star is demonstration technology, convention attendees did not talk to SGT Star directly, but passed their questions to a human handler who talked into the microphone. In order to get the best speech recogni-

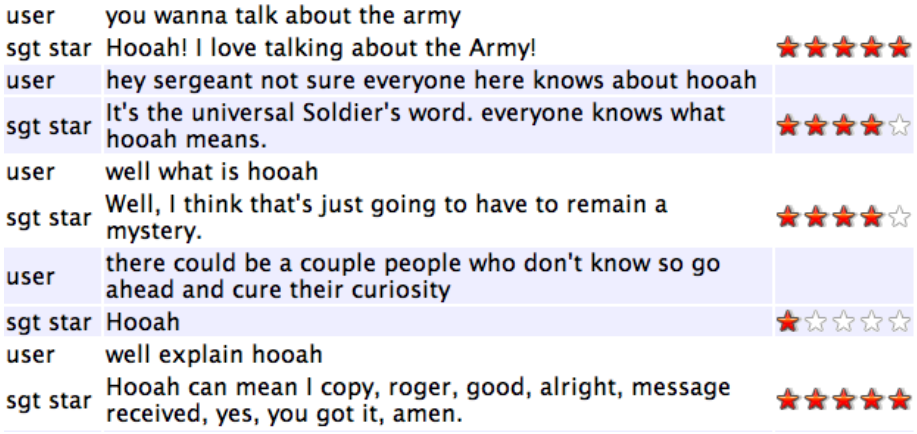

Figure 3: The rating interface

tion in a noisy convention environment, the acoustic models were tuned to the three individuals who did most of the handling.

To motivate convention attendees to explore the Army exhibit, the Army devised a task which required attendees to gather specific pieces of information from various parts of the exhibit in order to win a prize; the information they had to get from SGT Star was the meaning of "hooah" (a U.S. Army expression) and his Military Occupational Specialty (MOS). As a result, many of the dialogues from the convention are extremely short, and a disproportionate number of questions ask about "hooah" and SGT Star's MOS (approximately 17% and 13%, respectively).

3. Rating study

SGT Star's mission is to generate interest in learning about the Army and possible careers in it, but we have no way to measure the amount of interest generated. We assessed SGT Star's *coherence*, that is the appropriateness of his responses; the idea is that the more coherent a character is, the better he can engage the audience and create interest. An appropriate response to a question does not have to be a direct answer: a question or off-topic comment may sometimes be more appropriate, and SGT Star's off-topic responses were designed to allow him to hold a coherent conversation when he doesn't have a straight answer. We conducted a rating study in order to identify where SGT Star's coherence could be improved, to make him a more believable and engaging character.

SGT Star's performance resulted in a total of 3216 responses, and our study judged the appropriateness of these responses in context. The user utterances were transcribed individually, and entire dialogues (user utterances and SGT Star's responses) were presented as web pages on which judges rated each of SGT Star's responses on a scale of 1 to 5 (Figure 3). In 703 cases, the transcribed user utterance was identical to a training question and the response was linked to that question, and these were automatically rated as 5; the remaining 2513 responses were rated by the judges.

To ensure the ratings were meaningful we calculated inter-rater reliability using α (Krippendorff, 1980).[1] Three judges rated all 2513 responses, and a fourth judge (the first author) rated 474 of these. Overall reliability for the four judges was $\alpha = 0.789$; reliability for sub-groups of judges ranged from $\alpha = 0.901$ for the most concordant pair of judges to $\alpha = 0.676$ for the most discordant pair. Since overall reliability was close to the accepted threshold of 0.800, we continued the analysis by assigning each response the mean of all available ratings. Broken down by response type, reliability was high for on-topic responses ($\alpha = 0.794$) but barely better than chance for off-topic responses ($\alpha = 0.097$).

4. Response ratings

SGT Star has a total of 152 possible responses, of which 22 are tagged as off-topic. Off-topic responses are intended to be suitable both for genuine out-of-domain questions, for which SGT Star does not have

[1] Krippendorff's α is a chance-corrected agreement coefficient, similar to the more familiar K statistic (Siegel and Castellan, 1988). Like K, α ranges from -1 to 1, where 1 signifies perfect agreement, 0 obtains when agreement is at chance level, and negative values show systematic disagreement. The main difference between α and K is that α takes into account the magnitudes of the individual disagreements, whereas K treats all disagreements as equivalent; α is more appropriate for our study because the ratings are numerical, and the disagreement between ratings of 2 and 3, for example, is clearly lower than between 2 and 5. For additional background, definitions and discussion of agreement coefficients, see Artstein and Poesio (to appear).

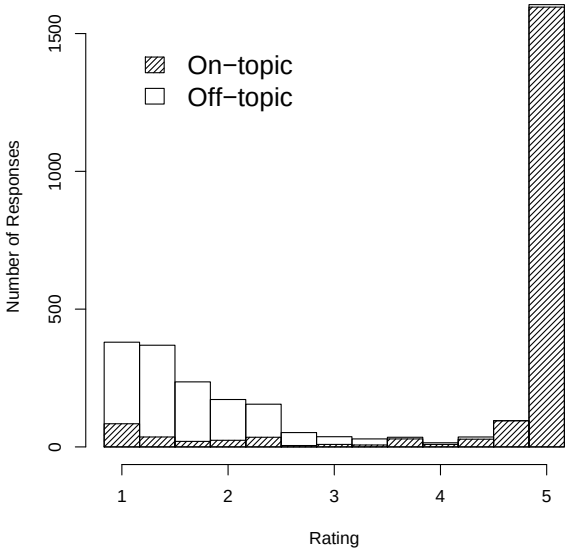

Figure 4: On-topic and off-topic ratings

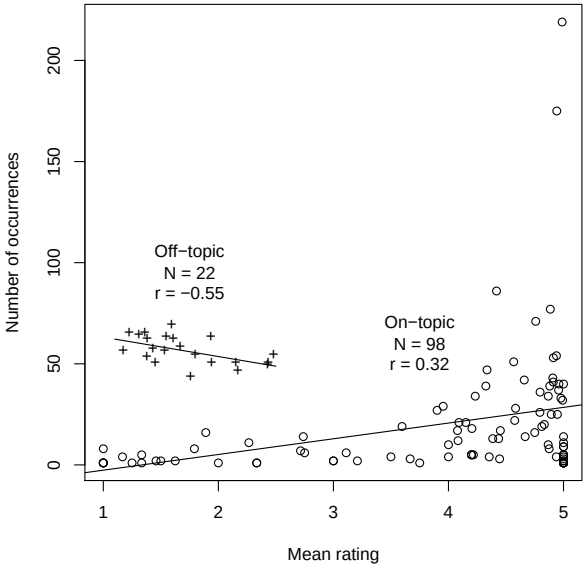

Figure 5: Rating and frequency correlations

an appropriate on-topic response, as well as for classifier failures due to factors like speech recognition errors or insufficient training data. The handlers at the convention were very familiar with SGT Star's range of responses and as a consequence there were very few out-of-domain questions; the vast majority of off-topic responses were a result of classifier failure.

The different responses were not all used to the same extent: in the testing, SGT Star produced 120 different responses (including all 22 off-topics), and their distribution was not even. This skewing is due to the uneven distribution of questions: The two most frequent responses by far, used 175 and 219 times, answer questions about "hooah" and SGT Star's MOS, brought about by the convention attendees' task.

The mean rating of SGT Star's responses was 3.47, but very few responses were close to the mean: most responses were either very good or very bad (first quartile 1.67, median 4.75). About 57% of the responses were rated above 3 and 43% below 3; this split roughly correlates with the difference between on-topic responses (61.5%), of which 80.7% received the maximum rating of 5, and off-topic responses (38.5%), of which 80.1% were rated 2 or less (Figure 4). There was also a clear separation in the frequency of individual responses. The off-topic responses were all used with similar frequency (ranging from 43 to 69) and received mean ratings of less than 2.5. In contrast, the low-rated on-topic responses appeared much less frequently (maximum frequency 16 for ratings under 3.5), while frequent on-topic responses were rated much higher (Figure 5). There is a positive correlation between rating and frequency for on-topic responses ($r = 0.32, p < 0.002, df = 96$),[2] whereas for off-topic responses the correlation is negative ($r = -0.55, p < 0.01, df = 20$).

The correlation between rating and frequency for on-topic responses remains robust even when we remove questions about the more common topics such as "hooah" and SGT Star's MOS. The reason is probably that the handlers quickly learned which responses were easy to elicit and popular with the crowd, and then targeted their questions to elicit these responses. The result was a selection of question topics narrower than SGT Star's full repertoire, which led to an overall good performance.

The negative correlation between rating and frequency for the off-topic responses was unexpected, since agreement on off-topics was low and individual off-topic responses are chosen at random. However, some off-topic responses are also linked to out-of-domain questions in the training data (for example, the response "ha ha, you're a bad man" is linked to the question "so do you have a girlfriend?"). The linked responses are expected to occur more frequently. As it turns out, requests for repetition ("I didn't hear that, could you repeat the question?") are usually not linked to any question, but these received higher ratings than the linked off-topic responses.

[2]The correlation is stronger if we use log frequencies: $r = 0.48, p < 0.001, df = 96$.

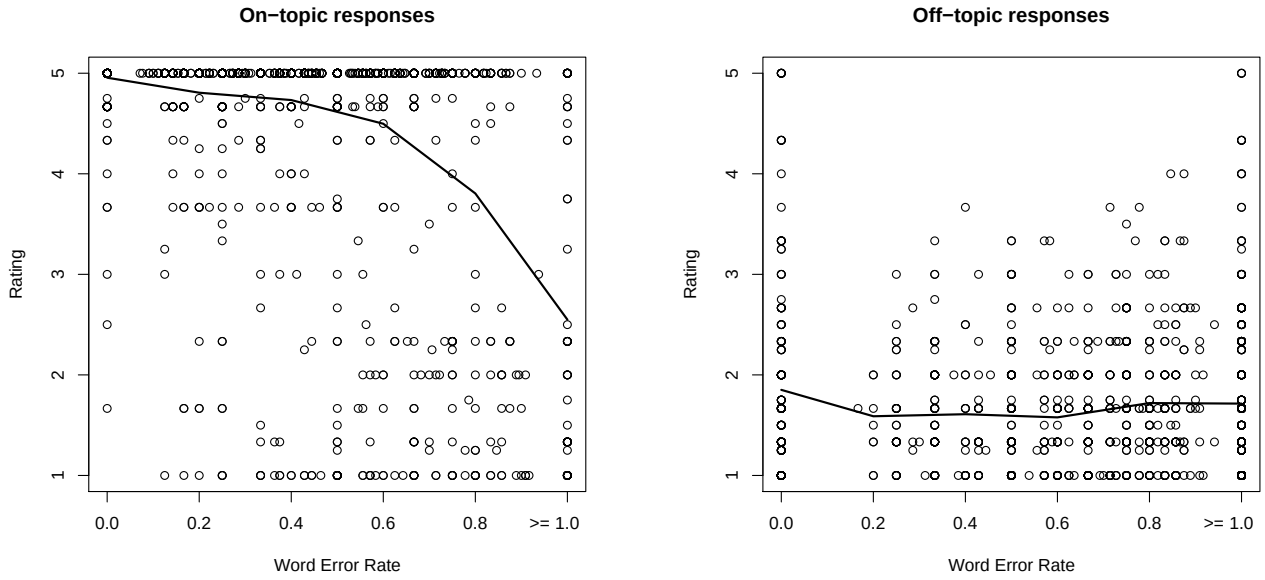

Figure 7: Word error rates and ratings: the lines show the mean rating for each WER band.

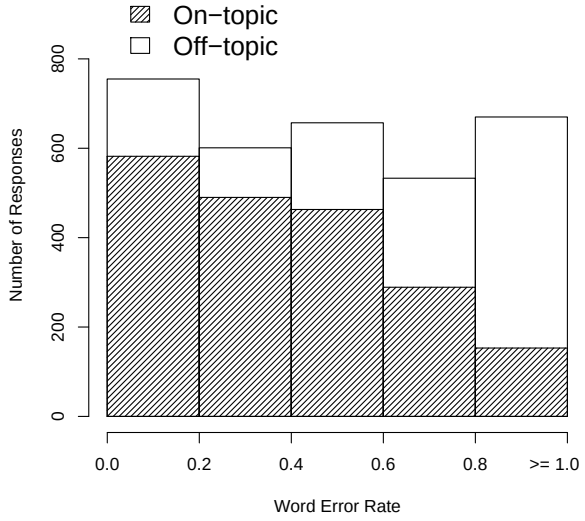

Figure 6: Word error rates

5. Speech recognition

Automatic speech recognition (ASR) affects performance (Leuski et al., 2006): if what SGT Star hears doesn't match what the user said, then SGT Star's response is more likely to be inappropriate. We computed the word error rate for each user utterance by comparing the ASR output with the transcribed speech.[3] Mean word error rate was 0.469, with an approximately uniform distribution; higher word error rates were more likely to trigger off-topic responses (Figure 6).

We found a negative correlation between the rating of SGT Star's response and the word error rate of the immediately preceding user utterance ($r = -0.47, p < 0.001, df = 3214$). This is partly due to the large block of off-topic responses with low ratings and high word error rates; however, the on-topic responses on their own also exhibit a (slightly weaker) negative correlation between response rating and word error rate ($r = -0.40, p < 0.001, df = 1975$). The off-topic responses do not show a similar correlation ($r = -0.02, p > 0.4, df = 1237$). The relations between response rating and word error rate of the preceding utterance are shown in Figure 7.

The negative correlation between rating and word error rate is expected: the less SGT Star understands the spoken utterance, the less likely he is to come up with a suitable on-topic response. Off-topic responses should not degrade with the mismatch between actual and recognized user utterance. One might even expect to find an improvement: due to the statistical language modeling in the ASR component, misrecognition of spoken words is more likely for out-of-domain questions, and SGT Star's off-topic responses should be more appropriate for those. We have not found this kind of effect, possibly because there were few out-of-domain questions.

[3]Word error rate is the number of substitutions, deletions and insertions needed to transform one string into the other, divided by the number of words in the actual (transcribed) speech; values above 1 were recorded as 1.

6. Conclusions

The rating study of data gathered in SGT Star's field deployment allowed us to study his functioning in

the situation for which he was designed, though with somewhat different parameters, namely being repeatedly asked for two pieces of information. The results show an interplay between SGT Star and his handlers, who are working to help the virtual character give his best performance. It is clear that SGT Star would have performed very differently if arbitrary users were allowed to ask unrestricted questions; dealing with such users and out-of-domain questions is the focus of another study, SGT Blackwell (Robinson et al., 2008). The study confirmed that speech recognition is a major obstacle – this is a difficult problem in the noisy environment where SGT Star operates. The study also identified off-topic responses as a place with substantial room for improvement, perhaps along the lines of Patel et al. (2006).

The rating study combined data extracted from system logs (ASR results and SGT Star's responses) with manual transcription, a human rating study, statistical testing and qualitative assessment. A question that comes up naturally is whether this method of evaluation can be automated or made less human-intensive. There is definitely some room for saving – for example, once we have established that the ratings are reliable, it is sufficient to have just one judge rate each response. However, rating the responses is not where most of the human effort went. All user utterances need to be manually transcribed, because the appropriateness of responses needs to be judged relative to the actual user utterance (this manual transcription is independently needed in order to improve performance of the highly domain-specific speech recognition models). But the most labor-intensive part is probably the analysis of individual responses. This is because we are not merely interested in a score that reports SGT Star's performance, but are also seeking to improve it for future exhibits. SGT Star's ability to respond appropriately depends on his training data, which consist of a list of questions, a list of responses, and links between the two. The questions come from actual user data, the responses reflect what we want SGT Star to be able to talk about, and the links come from a careful analysis of appropriateness which can only be achieved by manually examining actual conversation transcripts.

7. Acknowledgments

SGT Star is loosely based on a character created by Next IT for the Army's recruiting web site.[4] Thanks to two anonymous reviewers, to Kip Haynes for taking SGT Star on the road, and to Jacob Cannon, Jillian Gerten and Joe Henderer for rating SGT Star's utterances. The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

[4] http://www.GoArmy.com/

8. References

Ron Artstein and Massimo Poesio. to appear. Inter-coder agreement for computational linguistics. *Computational Linguistics*. Pre-publication draft at http://cswww.essex.ac.uk/Research/nle/arrau/icagr-short.pdf.

Klaus Krippendorff, 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12, pages 129–154. Sage, Beverly Hills, CA.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27, Sydney, Australia, July. Association for Computational Linguistics.

Ronakkumar Patel, Anton Leuski, and David Traum. 2006. Dealing with out of domain questions in virtual characters. In Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier, editors, *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Artificial Intelligence*, pages 121–131. Springer, Berlin.

Susan Robinson, David Traum, Midhun Ittycheriah, and Joe Henderer. 2008. What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. In *LREC 2008 Proceedings*, Marrakech, Morocco, May.

Sidney Siegel and N. John Castellan, Jr, 1988. *Nonparametric Statistics for the Behavioral Sciences*, chapter 9.8, pages 284–291. McGraw-Hill, New York, second edition.

Author Index