

Evaluation of HLT-tools for less spoken languages

Cristina Vertan, University of Hamburg, Germany,

vertan@informatik.uni-hamburg.de

Europe offers an unique context, not only due to the variety of spoken languages but also due to the necessity to offer a huge amount of information at least in all official U-languages.

During last years the number of mono- and multilingual systems dealing with other languages than the most frequent used ones (EN, FR, DE) increased. However most part of these systems are either developed for a particular scenario or tuned on specific (available) corpora. Consequently it is very difficult to evaluate such systems, as there are no (or less) reference test data and no (or very few) reference evaluation scores.

The evaluation of systems dealing with languages not so frequently spoken is a real challenge, which has to be urgently considered:

1. to ensure the development of qualitative similar systems across languages
2. to encourage researchers to develop tools for other languages than the frequently spoken ones.

Due to the big amount of languages, and their possible combinations one cannot build reference systems for all types of NLP tasks and all language combinations

On the other hand following actions are a viable solution for pushing forwards the evaluation and thus development of systems dealing with less spoken languages:

- ensure a reference parallel corpora
- build reference test suites
- define reference measures for classical NLP Applications (MT, IR, CLIR, QA, etc.) – underway.
- develop when possible language independent system models (like Moses in MT)

For the moment the only existent corpus covering 22 languages in the EU is the JRC-Acquis Multilingual Parallel Corpus (<http://langtech.jrc.it/JRC-Acquis.html>). The main problem with this corpus is the special type of language it covers, namely law. Systems trained on this corpus will give less convincing results on other type of texts. The other way around, systems trained on different texts will deliver suboptimal results on JRC-Acquis. Similar problems can be observed on other corpora as the OPUS Corpus (<http://urd.let.rug.nl/tiedeman/OPUS/>). Some other corpora developed in frame of earlier EU-Projects like MULTTEXT-EAST (<http://nl.ijs.si/ME/CD/docs/mte-corp.html>) or n-ouse projects like ROGER (Romanian - German - English - Russian corpus, <http://nats-www.informatik.uni-hamburg.de>) are either encoded in old fashioned standards, or unknown to the large research community, and themselves too small to be used for large-scale applications

We propose the building of a parallel corpus for all European languages, covering different domains: law, medicine, news, tourism. This should be a test-bed for any monolingual and crosslingual application. Steps for building such corpus are:

1. Unify annotation for existent parallel corpora: JRC-Acquis, OPUS, Europarl, MULTEX., MULTEX-EAST
2. Investigate all EU projects with participation from various language communities Networks of Excellence. Especially in non-technical domains project reports are translated in all participant languages (e.g. CALIMERA <http://www.calimera.org/default.aspx>, I*Teach <http://i-teach.fmi.uni-sofia.bg/>)
3. Investigate existent translations of the official EU Web sites as at least part of the information is made available in all languages
4. Collect parallel corpora produced in different projects

In a second steps reference measures have to be defines for each type of application. These measures should consider the data-set profiling aspect (see DeRoeck&Al., Invited Talk at RANLP 2005) as languages do not have the same distribution. Experience in CLEF Competitions should be a good starting point.

Special attention requires Machine Translation. The recent development of the Moses open source system, offers the possibility to develop baseline Systems for any language pair, under the assumption that a significant amount of training data (parallel corpora exist).

In conclusion a possible roadmap for bringing forwards the evaluation of systems working with less spoken languages in Europe relies on following phases:

1. Define a standard for minimal parallel corpora annotation (to be used as input data)
2. Development of a reference parallel corpus in all (most part) of the accepted EU languages.
3. Releasing of test suites (parts of this corpus) to be used in test scenarios.
4. Define standard test scenarios for evaluating (LT-systems) on these test-suites.
5. Publish A dataset profiling study, considering influence of language distribution on the statistical mechanisms for all official EU-languages and major HLT applications
6. Implementation of baseline MT-systems using Moses and testing on the reference corpus