Improving Machine Translation Performance Using Comparable Corpora

Andreas Eisele, Jia Xu

DFKI GmbH

{Andreas.Eisele,Jia.Xu}@dfki.de

3rd Workshop on Building and Using Comparable Corpora
Valletta, Malta
2010-05-21

Talk Overview

- The need for multilingual NLP resources
- Statistical MT as a step towards larger goals
- Problems with sparse data and ways ahead
- Concrete next steps
- Relation to other projects
- Conclusions

The Need for Multilingual NLP

- Despite impressive results, work on natural language processing has focussed on a small number of languages, mainly English
- Most EU citizens need such technology in their mother language, e.g. MT from "big" to "small" languages
- Focus on morphologically simple languages like English has also lead to relative weaknesses in the treatment of richer morphologies in the current state of the art
- High-quality MT (and NLP in general) needs to be based on a combination of **linguistic knowledge**, generally from grammars and rules, with **extra-linguistic knowledge** found in text corpora
- EuroMatrix Plus investigates hybrid approaches to MT

Types of Relevant Knowledge

We need knowledge sources of many different types

- Linguistic knowledge
  - Mappings from words to parts of speech
  - Morphological regularities
  - Lemmatization
  - Compounds and agglomerative constructions
  - Linguistic features (case, number, gender, tenses, ...)
  - Dependencies between words and constituents
  - Semanctic roles and relations
- Cross-lingual knowledge on several levels
  - Lexical and terminological correspondences
  - Structural correspondences between languages
  - Correspondences on level of features
- Extra-linguistic knowledge found in text
  - Patterns of typical usage
  - World knowledge

Knowledge acquisition bottleneck

- Recent progress in many areas shows that important knowledge can be derived from text corpora
- Supervised machine learning works well, but...
  - requires expensive annotation of data
  - leads to domain-specific models
  - not feasible for 20+ languages across many domains
- Training of statistical MT models is a way to induce knowledge from real-world data, using translation as a replacement for annotation
- We can learn cross-lingual correspondences, but...
  - Strong dependency on parallel corpora
  - Induction of language-specific knowledge requires mixed approaches

Overcoming the acquisition bottleneck

... via bootstrapping:

- Use small parallel corpora, existing lexicons, terminologies, and MT engines to
  - build partial cross-lingual models
  - map linguistic annotations into corpora of new languages
  - derive approximations of linguistic annotations and tools for these languages
- Use such approximations to find cross-linguistic correspondences even in non-parallel corpora
- Increase coverage via interative application
- Keep accuracy high via manual inspection of conflicting results

A closer look on SMT training

- SMT training tries to explain text in one language given a corresponding text in some other language
- Typical reasoning step:

  Assume we know A B C ⇔ X Y Z, A⇔Z, B⇔Y

  Conclude that C ⇔ X
- But in real life:
  - A ⇔ Z, ... are themselves only guesses from the data
  - Translations in parallel corpora are not always very close

➔ SMT training needs to cope with mismatches and inaccuracies

- SMT training (e.g. GIZA++) performs bootstrapping of knowledge from uncertain/risky assumptions
- Initial high error rates decrease, as errors tend to spread randomly over many different hypotheses, whereas the true facts accumulate higher frequency counts ➔ more data leads to better separation between signal and noise

Parallel vs. comparable corpora

- The distinction is actually not quite clear-cut, rather gradual, e.g. many phrase pairs within EuroParl are not mutual translations
- Techniques for locating parallel bits in comparable corpora have been presented since many years
- Better control of usage of risky assumptions in SMT training can increase expected performance of these techniques on comparable corpora
- More linguistic features help to increase alignment quality (see e.g. several papers at this LREC)
- They might be indispensable for properly exploiting comparable corpora
- Fine-tuning the combination of multiple knowledge sources (linguistic, statistic) requires research effort

Initial Steps

- Collect large amounts of parallel and comparable corpora
  - Acquis Communautaire
  - TMs and corpora from technical domains
  - News corpora
  - Wikipedia articles
- Find parallel snippets in comparable corpora
  - Use bootstrapping as sketched on earlier slides
- Use extracted data to build SMT models
- Estimate accuracy for phrase pairs obtained from comparable corpora by counting samples
- Use such estimates within SMT decoding, giving priority to clear cases
- Optimize relative weights of different knowledge sources via MERT techniques

MERT optimization for combining knowledge source

- From LREC poster/upcoming EAMT paper:

  Use MERT to combine knowledge from different sources

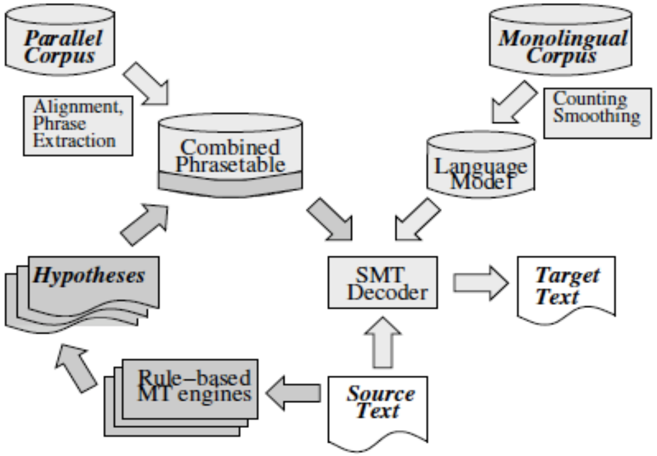
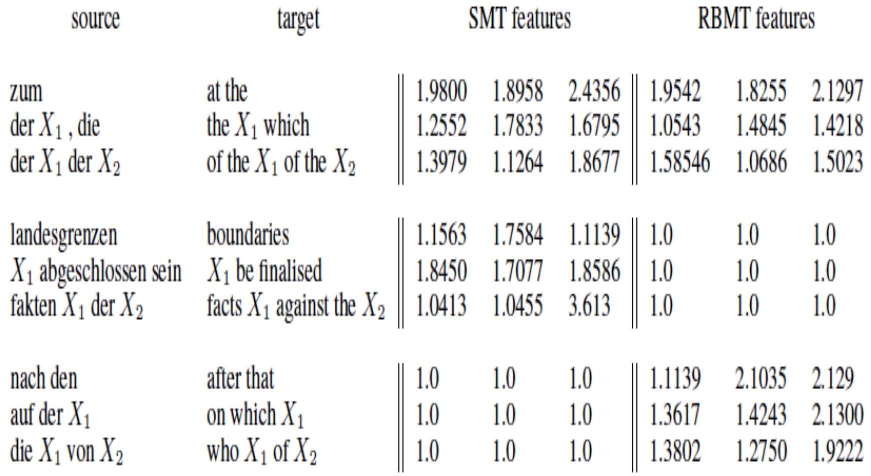


Figure 1: Hybrid architecture of the system



source	target	SMT features			RBMT features		
zum	at the	1.9800	1.8958	2.4356	1.9542	1.8255	2.1297
der X_1 , die	the X_1 which	1.2552	1.7833	1.6795	1.0543	1.4845	1.4218
der X_1 der X_2	of the X_1 of the X_2	1.3979	1.1264	1.8677	1.58546	1.0686	1.5023
landesgrenzen	boundaries	1.1563	1.7584	1.1139	1.0	1.0	1.0
X_1 abgeschlossen sein	X_1 be finalised	1.8450	1.7077	1.8586	1.0	1.0	1.0
fakten X_1 der X_2	facts X_1 against the X_2	1.0413	1.0455	3.613	1.0	1.0	1.0
nach den	after that	1.0	1.0	1.0	1.1139	2.1035	2.129
auf der X_1	on which X_1	1.0	1.0	1.0	1.3617	1.4243	2.1300
die X_1 von X_2	who X_1 of X_2	1.0	1.0	1.0	1.3802	1.2750	1.9222

Figure 2: Example entries from combined phrase table

- Variants of this approach can be used to combine phrase pairs from different types of corpora, e.g. to combine “parallel” with “comparable” material

Next Steps

- Use first generation of SMT models in a bootstrapping loop, try to improve accuracy of extraction from comparable corpora
- MERT optimizing BLEU scores may not be ideal; we need to explore alternative scoring methods
- Incorporate distinction between parallel and comparable sources into alignment algorithms

  Similar to semi-supervised alignment techniques combining annotated with un-annotated data, we can combine parallel with comparable corpus data
- Induce linguistic features such as PoS classes via cross-lingual projection and use them to improve alignment

Synergies between projects

- EuroMatrix Plus builds (among many other things)
  - Statistical and hybrid MT models for EU language pairs
  - Infrastructure for making MT engines available and collecting feedback (WikiTrans)
  - Advanced leaning methods (including work on comparable corpora!)
  - Methods for improving models through feedback
- Many of these modules can be adapted to the work with comparable corpora
  - Baseline SMT models can be used for identifying parallel pieces in comparable corpora
  - Feedback on MT results reveals insights on pros/cons of baseline SMT vs. SMT from comparable data
  - Methods for model update can be adapted to obtain sharper distinction between signal and noise

Conclusion

- Techniques for knowledge extraction from parallel text can be generalized to comparable corpora
- Methods for training and using SMT can be adapted to and optimized for the generalized setting
- ACCURAT and EuroMatrix Plus complement in the methods they apply
- They also complement each other in the coverage of language pairs
- High-quality MT will need to combine corpus-based evidence with many types of linguistic knowledge,
- hence these approaches should be seen as steps on a longer path towards the construction of linguistically informed approaches to NLP and MT for a large subset of European languages

Thank you for your attention.

www.euromatrixplus.eu

www.accurat-project.eu

The research within the projects EuroMatrix Plus and Accurat has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreements nº 231720 and nº 248347.