Experiments with String Similarity Measures in the EBMT Framework

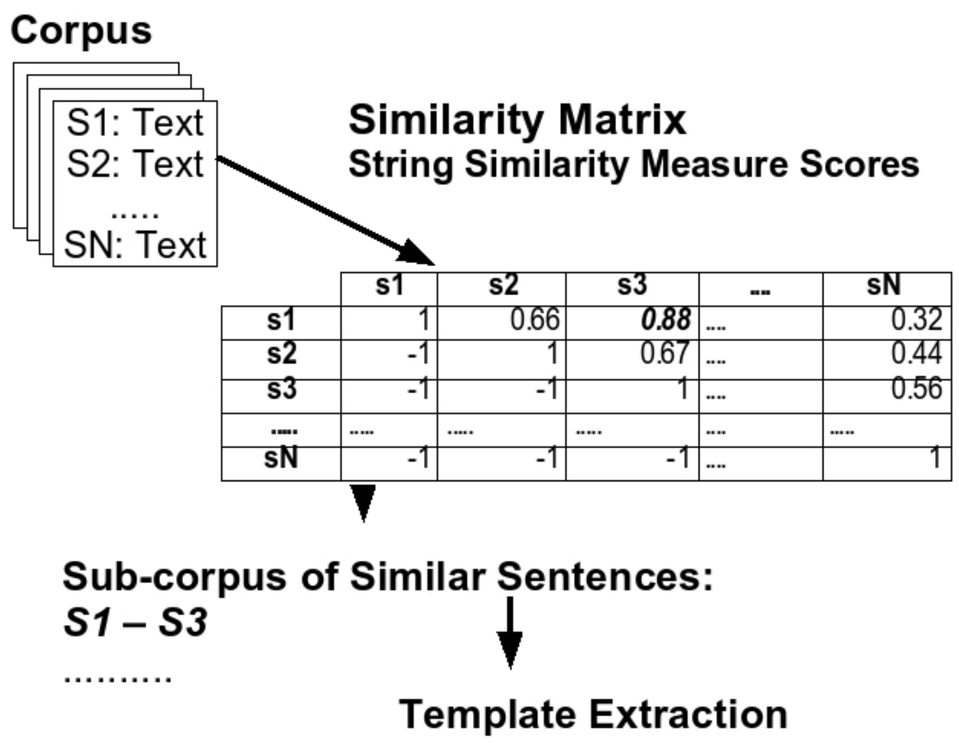
Natalia Elita, Monica Gavrila, Cristina Vertan
Natural Language Systems Group, University of Hamburg
{elita, gavrila, vertan}@informatik.uni-hamburg.de

Goal:
Retrieving similar sentences as candidates for translation templates to be used in an EBMT system.

Approach:
Similarity Matrix; analysis of 21 string similarity measures (character-, token-based, hybrid).

Advantages:
The search space for templates is considerably reduced.

Results:
Token-based measures (Overlap Coefficient with Common Words) are the most suitable for translation template extraction.



Sentence Similarity Constraints

Two sentences are similar, if:

1. they have at least three words in common (Common Words - CW - Threshold);

2. the sequence of common elements consists of at least 50% content words (lexical words);

3. one sentence is a sub-sentence of the other to the proportion of 50% (Overlap Coefficient - OC - Threshold).

$$OC(s1, s2) = \frac{(|s1\&s2|)}{min(|s1|, |s2|)},$$

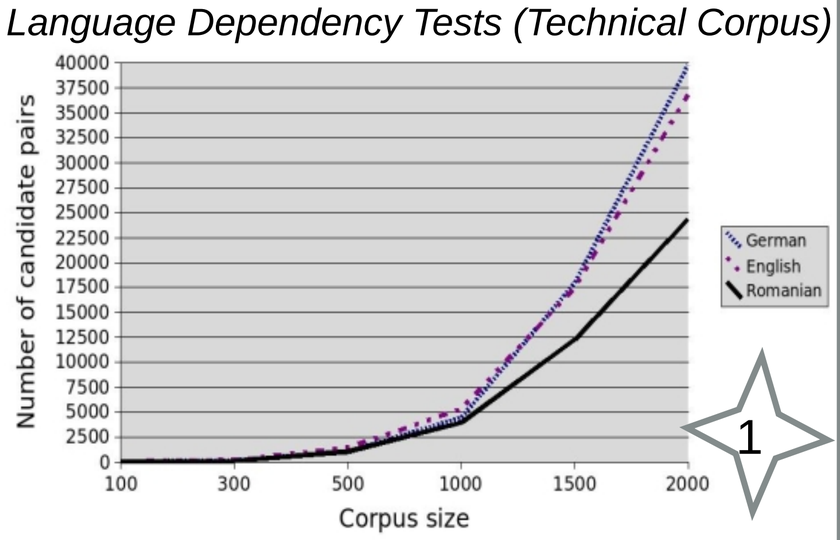
where |s| is the number of tokens in s, and |s1&s2| the number of common tokens in s1 and s2.

Example

Settings for text and e-mail messages
Settings for the multimedia messages

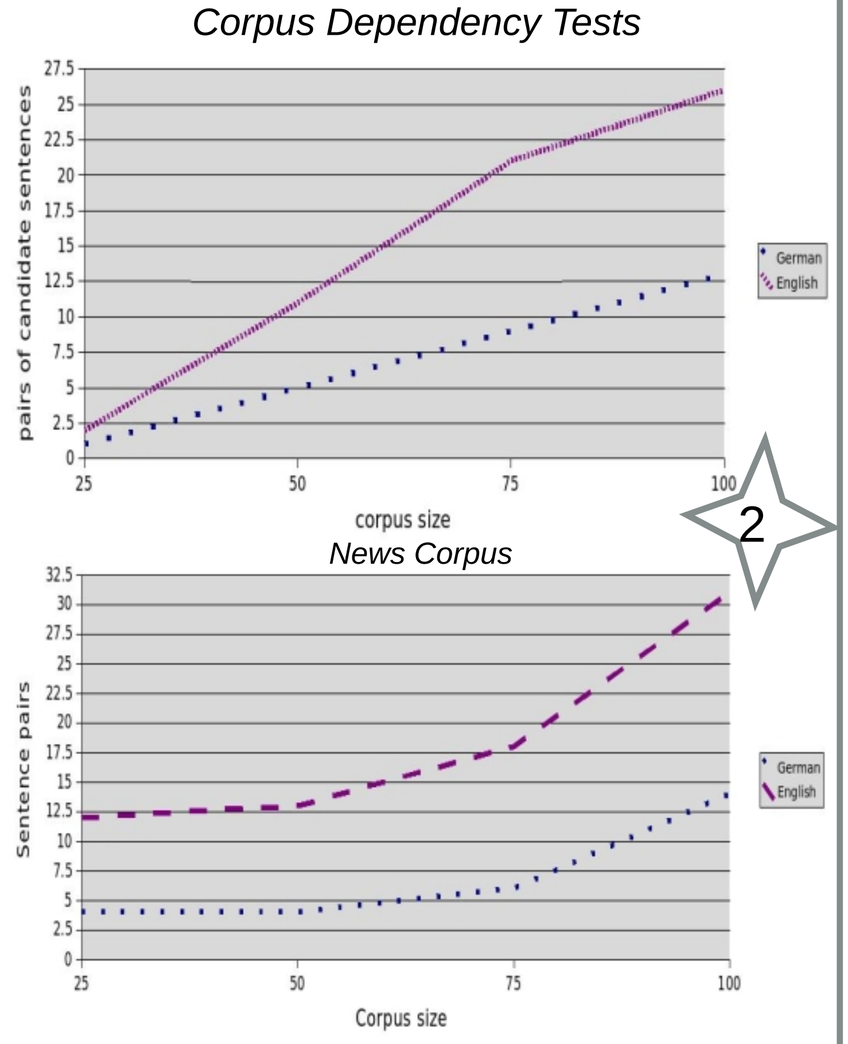
Experiments

Language Dependency Tests (Technical Corpus)



1

Corpus Dependency Tests



2

Technical Corpus

1 Smaller number of sentences extracted for Romanian, due to the following reasons
- it belongs to a different language family;
- it is highly inflected.

2 The number of extracted sentence-pairs is corpus dependent. More candidates are extracted for the technical corpus because it is domain restricted.

Conclusions

- Similarity measures' thresholds are language sensitive.
- The number of extracted sentence-pairs is language / corpus dependent.
- Token-based measures give best results for our goal.

References

* *SimMetrics. http://www.dcs.shef.ac.uk/sam/simmetrics.html.*

http://nats-www.informatik.uni-hamburg.de/view/User/PhD