Estimating language relationships from a parallel corpus. A study of the Europarl corpus

Taraka Rama, Lars Borin

Språkbanken
Department of Swedish Language
University of Gothenburg

NODALIDA 2011

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

Taraka Rama
and Lars Borin

Introduction
Previous Work
Our Approach
Dataset
Experiments
Results
Final Words
References

Outline

Introduction

Previous Work

Our Approach

Dataset

Experiments

Results

Final Words

References

Lexicostatistics I

Objective is to automatically identify the genetic relationships between languages from parallel corpus.

- Estimate the distance matrix between the languages.
- Use a clustering algorithm to infer the family tree for the languages.

Lexicostatistics II

Some concepts.

- Cognates are words which are genetically related. Ex: English/German: **hound/Hund**; English/Armenian: **two/erku**.
- Loanwords from other languages are not considered as cognates.

Taraka Rama and Lars Borin

Lexicostatistics III

Some assumptions about using lexical items for estimating the distance matrix.

- A word list of length of 40-200 basic meanings is collected for every language.
- Expert cognacy judgements are made between the word pairs in the lists.
- Expert judgement is based on comparative method.
- Cognates are identified using recurrent sound correspondences.
- The number of cognates between the two languages is judged as the similarity between the two languages.
- These steps is known as *lexicostatistics* in historical linguistics.

Lexicostatistics IV

Automatic Identification of cognates.

- Orthographic measures are generally used for judging the similarity between the word pairs.
- Methods such as HMMs require initial training data.
- Computational linguistics use the term *cognates* in a broader sense.
- Includes loanwords and chance resemblances, *false positives*.
- No way of identifying genetically related but having different forms, *false negatives*.

Outline

Introduction

Previous Work

Our Approach

Dataset

Experiments

Results

Final Words

References

Taraka Rama
and Lars Borin

Introduction

Previous Work

Our Approach

Dataset

Experiments

Results

Final Words

References

The previous work which comes closest to the work presented here is that of (Koehn 2005), who trains pair-wise statistical translation systems for the 11 languages of the Europarl corpus and uses the systems' BLEU scores for clustering the languages, under the assumption that ease of translation correlates with genetic closeness.

Outline

Introduction

Previous Work

Our Approach

Dataset

Experiments

Results

Final Words

References

Automatic Identification of Cognates I

- Automatically identify word pairs which are translations of each other.
- Use a orthographic measure for computing the similarity between each word pair.
- Remove word pairs which are below a particular *cutoff*.
- We use *longest common subsequence ratio* as the orthographic measure.
- The *cutoff* is fixed at 0.58 to account for length bias.

Automatic Identification of Cognates II

Given the cognate lists for two languages, the distance between two languages l_a, l_b can be expressed using the following equation:

$$Dist(l_a, l_b) = 1 - \frac{\sum_i sim(l_a^i, l_b^i)}{N} \tag{1}$$

$sim(l_a^i, l_b^i)$ is the similarity between the *i*th cognate pair and is in the range of [0, 1].
N is the number of words being compared.

Automatic Identification of Cognates III

String similarities is only one of the many possible ways for computing the similarity between two words.

Lexicostatistics is a special case of above equation where the range of *sim* function is $0|1$.

Definitions:

- Levenshtein distance is defined as the minimum number of insertions, deletions and subtractions to transform a string into other.
- Dice is defined as twice the overlap of the number of bigrams divided by the total number of bigrams.
- LCSR is defined as the length of the longest common subsequence divided by the maximum length of the two strings.

Outline

Introduction

Previous Work

Our Approach

Dataset

Experiments

Results

Final Words

References

Europarl Corpus

- The publicly available Europarl corpus was used.
- The corpus is from English to ten languages.
- 45 pairs of parallel corpora were created by using English as bridge language.
- The first 100,000 sentences were included.
- Every language except Finnish is a Indo-European language.
- All the other languages fall into different branches of Indo-European language family, Germanic and Romance.

Outline

- Introduction
- Previous Work
- Our Approach
- Dataset
- Experiments
- Results
- Final Words
- References

- The freely available statistical machine translation system MOSES (Koehn et al. 2007) was used for aligning the words.
- Word alignments were used for extracting the cognate pairs.
- For every language pair, word pairs with **LCSR** less than cutoff were removed.
- We experimented with three string similarity measures Levenshtein Distance, Dice and LCSR.
- All the measures are symmetric.
- UPGMA as implemented in PHYLIP was used to cluster these distances.

Outline

Introduction

Previous Work

Our Approach

Dataset

Experiments

Results

Final Words

References

Language	# Probable Cognates
English	1458
German	1043
Dutch	1489
Swedish	**2624**
Danish	2149
French	955
Spanish	823
Portugese	831
Italian	1333

Table: The number of probable cognates of each language with Finnish.

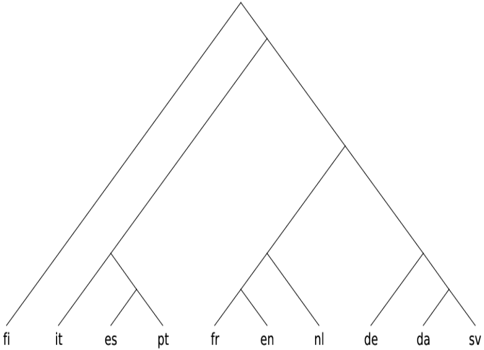

Figure: **Levenshtein Distance**: UPGMA tree.

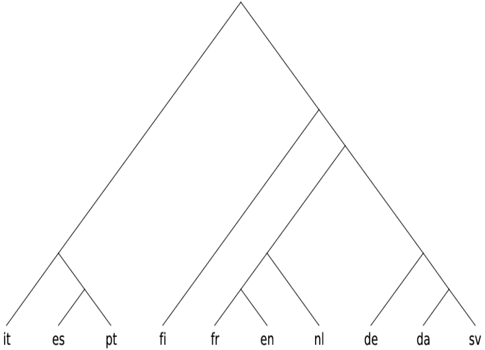

Figure: **Dice**: UPGMA tree.

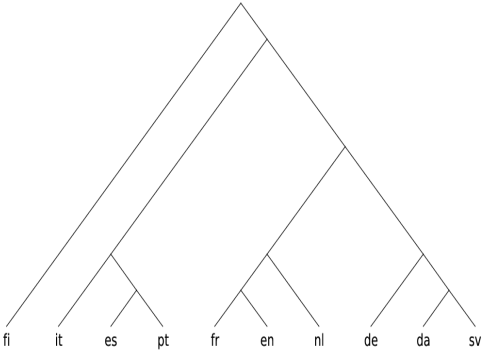

Figure: **LCSR**: UPGMA tree.

Discussion

- Finnish shares the highest number of cognates with Swedish.
- Working with corpus avoids the subjectivity involved in collecting the Swadesh list.
- It also brings in automation which is not available in (**?**).
- The tree on the whole agrees with the commonly accepted subgrouping.
- Comparing with (Koehn 2005) it returns lower order relationships better than the higher order.

Outline

Introduction

Previous Work

Our Approach

Dataset

Experiments

Results

Final Words

References

Conclusion

- The preliminary results indicate that a parallel corpus could be used for this kind of study.
- Dutch, English and French might have borrowed large parts of the vocabulary used in the Europarl corpus (administrative and legal terms) from French, and additionally in many cases have a spelling close to the original French form of the words.

- Automatically distinguish cognates from loanwords.
- Incorporate syntactic and semantic features in the future.
- Use POS tags and context vectors for estimating the similarity.

Outline

Introduction

Previous Work

Our Approach

Dataset

Experiments

Results

Final Words

References

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama and Lars Borin

References

Koehn, P. (2005), Europarl: A parallel corpus for statistical machine translation, *in* ‘MT summit’, Vol. 5.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. et al. (2007), Moses: Open source toolkit for statistical machine translation, *in* ‘Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions’, Association for Computational Linguistics, pp. 177–180.