

# Statistical Translationese

## Learning Time-Aligned Orderings of Documents by Example of Historical Documents

Zahurul Islam

Text Technology Group, Goethe-Universität Frankfurt am Main

Monday, 4 July 2011

# Outline

# The problem



- The old Georgian gospel is translated from the old Greek gospel or the old Armenian gospel.
- Task: Find out the correct source of the old Georgian gospel.
- Why??



# The problem

- The old Georgian gospel is translated from the old Greek gospel or the old Armenian gospel.
- Task: Find out the correct source of the old Georgian gospel.
- Why??



# Example

## Translation

This here is one of the biggest icebergs in the long changeful history of seafaring.

## Candidate sources

1. Voici, c'est un des icebergs les plus grands dans l'histoire longue et variée de la marine.
2. Das hier ist einer der größten Eisberge in der langen wechselvollen Geschichte der Seefahrt.

courtesy: Armin Hoenen



# Example

## Translation

This here is one of the biggest icebergs in the long changeful history of seafaring.

## Candidate sources

1. Voici, c'est un des icebergs les plus grands dans l'histoire longue et variée de la marine.
2. Das hier ist einer der größten Eisberge in der langen wechselvollen Geschichte der Seefahrt.

courtesy: Armin Hoenen

# One more example



## Translation

Mr President, the Pope recently visited Vienna and said that we should be talking of Europeanization rather than enlargement towards the East.

## Candidate sources

1. Monsieur le Président, en visite à Vienne il y a peu, le pape a dit qu'il ne fallait pas tant parler d'élargissement à l'est que d'européanisation.
2. Herr Präsident! Der Papst war unlängst in Wien und hat gesagt, wir sollen nicht nur von einer Osterweiterung reden, sondern insbesondere von einer Europäisierung sprechen.



# One more example

## Translation

Mr President, the Pope recently visited Vienna and said that we should be talking of Europeanization rather than enlargement towards the East.

## Candidate sources

1. Monsieur le Président, en visite à Vienne il y a peu, le pape a dit qu'il ne fallait pas tant parler d'élargissement à l'est que d'européanisation.
2. Herr Präsident! Der Papst war unlängst in Wien und hat gesagt, wir sollen nicht nur von einer Osterweiterung reden, sondern insbesondere von einer Europäisierung sprechen.

# Target Corpora



Table: Statistics of the target corpora

| Corpus | Sentences | Tokens | Types |
|---|---|---|---|
| Old Georgian | 3,738 | 59,526 | 8,771 |
| Old Greek | 3,738 | 67,831 | 9,622 |
| Old Armenian | 3,738 | 67,225 | 9,537 |

# Challenges



## The problem

- The old Georgian gospel is translated from the old Greek gospel or the old Armenian gospel.
- Task: Find out the correct source of the old Georgian gospel.

## Challenges

1. Small amount of data
2. No other sources of knowledge

# Outline

# Nature of translated text (Baker 1996 and Hansen 2003)



- Simplification
  - Describes tendency of the translator to simplify text in order to make the text easier to read.
- Normalization
  - Typical patterns of the source language is visible in the translation
  - Example: Software manual
- Explicitation
  - Translation is less ambiguous than the source.
  - Uses explanatory vocabulary (e.g., therefore, consequently, that)



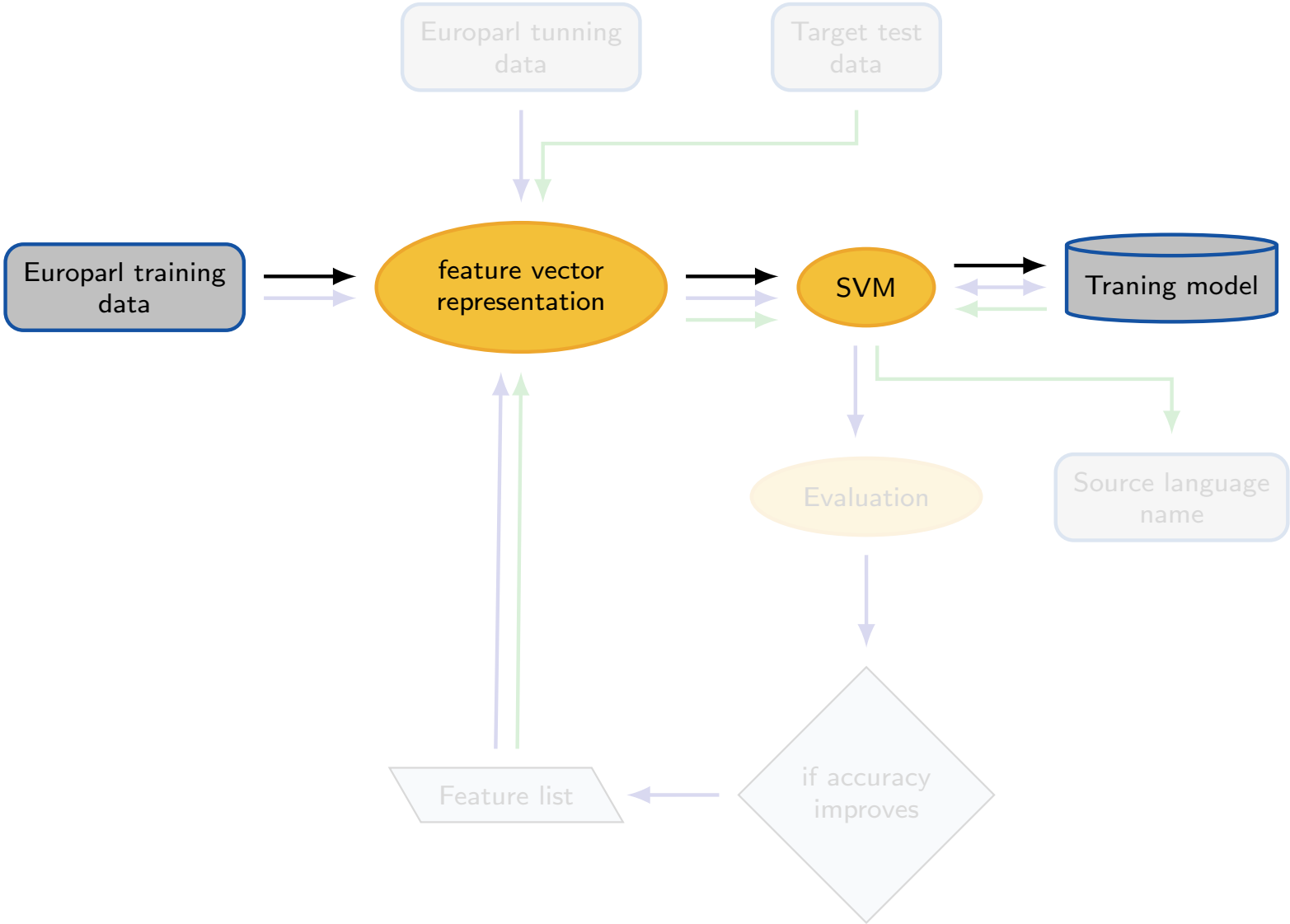
# Nature of translated text

- Readability (Pastor et al. 2008)
  - As the translated text is simpler then its source, the translation will be more readable
- Convergence (Pastor et al. 2008)
  - Translated texts tend to be more similar to each other than non-translated texts
- Collocational (Bernardini 2007)
  - Translated texts are more collocational than non-translated texts

# Outline

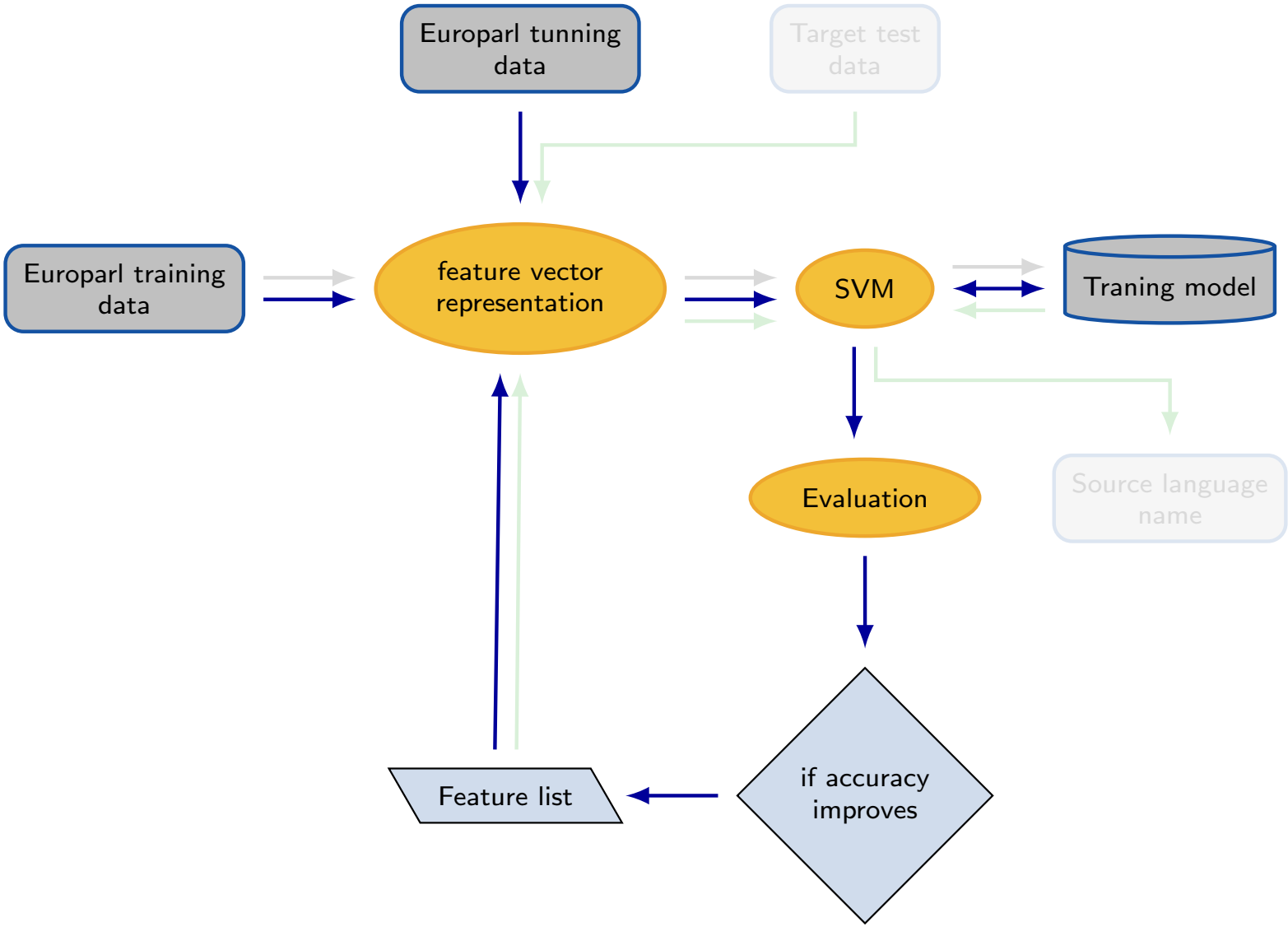


# Approach


Europarl tunning data
Target test data
Europarl training data
feature vector representation
SVM
Traning model
Evaluation
Source language name
if accuracy improves
Feature list

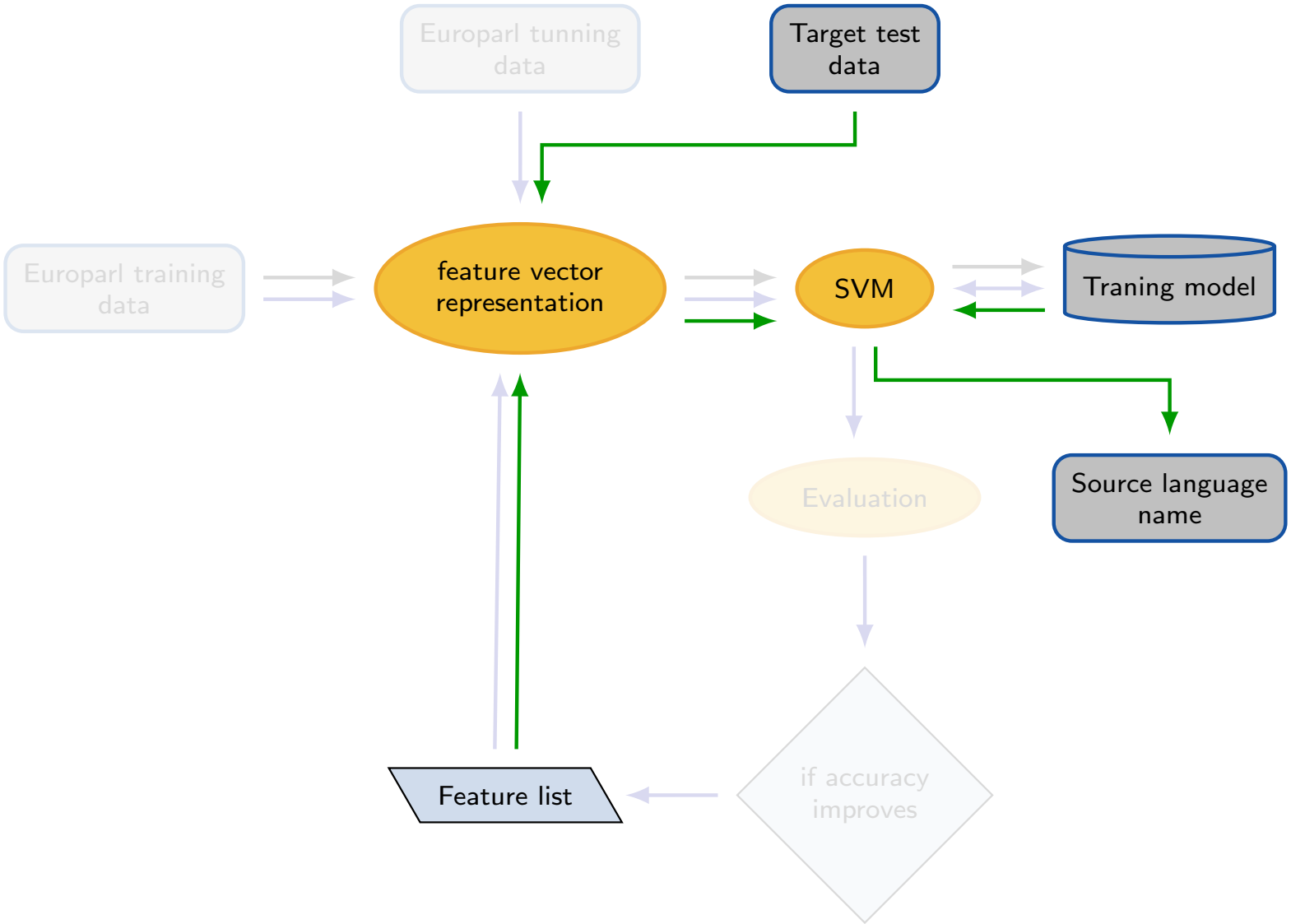


# Approach


Europarl tunning data
Target test data
Europarl training data
feature vector representation
SVM
Traning model
Evaluation
Source language name
if accuracy improves
Feature list

# Approach




Europarl tunning data
Target test data
Europarl training data
feature vector representation
SVM
Traning model
Evaluation
Source language name
if accuracy improves
Feature list



# Europarl corpora

- Version 6, released in February 2011
- Around 50 million words per language
- Marker
  1. Chapter: <CHAPTER id>
  2. Speaker and Language: <SPEAKER id name language>
  3. Paragraph: <P>



# Europarl preprocessing

1. Sentence alignment
2. Remove empty lines
3. Extract sources and their corresponding translations
4. Take the intersection
5. Tokenization
6. Lowercasing

# Europarl corpora statistics



Table: Number of sentences in the source

| | Sentences |
|---|---|
| German | 33,483 |
| English | 28,720 |
| French | 25,953 |
| Total | 88,156 |

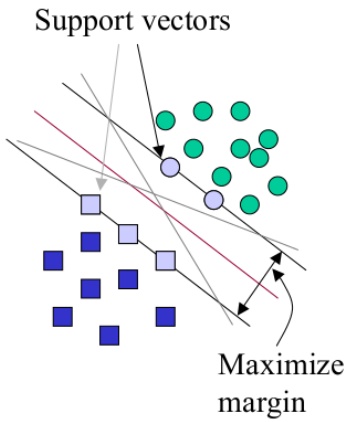
Table: Statistics of the training corpora

| Corpus | Tokens | Types |
|---|---|---|
| Source | 2,169,257 | 80,222 |
| Translation 1 | 2,238,958 | 76,117 |
| Translation 2 | 2,239,356 | 74,257 |

# Support vector machine



- Supervised learning method
- Support vectors are the data points that lie closest to the decision surface
- The classifier is a separating hyperplane
- Using SVM, we can map $X \rightarrow Y$ , where $x \in X$ is some data and $y \in Y$ is some class labels.




courtesy: R. Berwick



# Features

## Simplification (Hansen 2003 and Pastor et al. 2008)

1. Average sentence length
   - $\frac{\sum_{i=1}^{n} length(S_i)}{N}$
2. Lexical density
   - $\frac{\sum_{i=1}^{w} lexicalWord_i}{W}$
3. Type-token ratio
   - $\frac{\sum_{i=1}^{w} type_i}{W}$
4. Lexical richness
   - $\frac{\sum_{i=1}^{w} lemma_i}{W}$



# Features

## Explicitation (Hansen 2003 and Pastor et al. 2008)

1. Complex sentence ratio
   - $\frac{\sum_{i=1}^{n} complexSentence(S_i)}{N}$
   - Complex sentence: A sentence that contains more than one finite verb.
2. Number of explanatory words in a segment
   - Example: therefore, consequently, that and so on.



# Features

## Normalization (Baker 1996 and Hansen 2003)

- Hypothesis: Syntactic pattern of a translation is biased by the Source
- The way to capture this idea:
  1. Annotate training data with POS information
  2. Find the top ranked most frequent POS patterns in the translation and find their equivalence in the candidate sources
  3. Come up with a weighting function that will assign higher weights to higher ranked POS patterns than to lower ranked patterns
  4. Compute distribution of those patterns per segment



# Features

## Normalization (Baker 1996 and Hansen 2003)

- Hypothesis: Syntactic pattern of a translation is biased by the Source
- The way to capture this idea:
  1. Annotate training data with POS information
  2. Find the top ranked most frequent POS patterns in the translation and find their equivalence in the candidate sources
  3. Come up with a weighting function that will assign higher weights to higher ranked POS patterns than to lower ranked patterns
  4. Compute distribution of those patterns per segment



# Features

## Collocation (Bernardini 2007, Botafogo et al. 1992 and Mehler 2008)

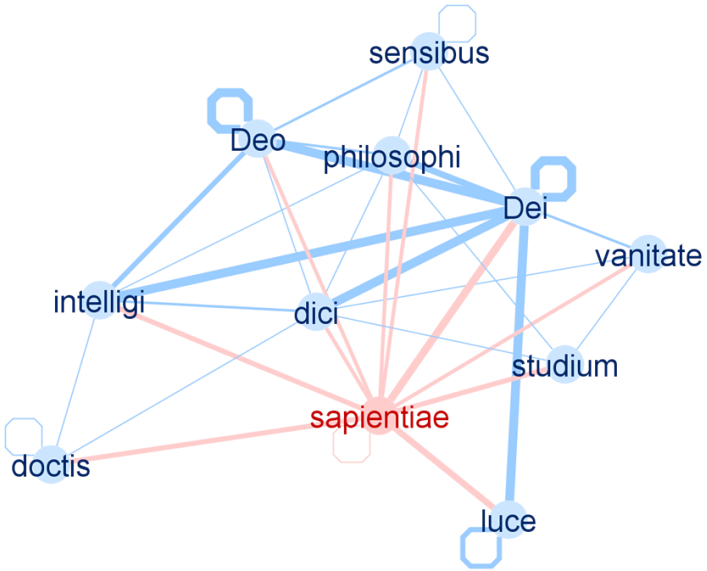
1. Number of significant $N$-grams
2. Undirected collocational graph: Compactness



# Features

## Compactness: $Cp(G)$ (Mehler 2008)

$$\frac{\left(Max(G)-\left(\sum_{v\in V}\sum_{w\in V} gd(v,w)+D_{max}(G)\sum_{G'\in Com(G)}|G'||V|-|G'|^2\right)\right)}{Max(G)-Min(G)} \in [0,1]$$




courtesy: Mehler et al. 2010



# Features

## Readability (Smith and Senter 1967, Coleman and Liau 1975, Pastor 2008)

1. Automatic Readability Index (ARI)

$$ARI = 4.71\left(\frac{chars}{words}\right) + 0.5\left(\frac{words}{sentences}\right) - 21.43$$

2. Coleman-Liau Index (CLI)

$$CLI = 5.89\left(\frac{chars}{words}\right) - 0.3\left(\frac{sentences}{words}\right) - 15.8$$

# One more problem

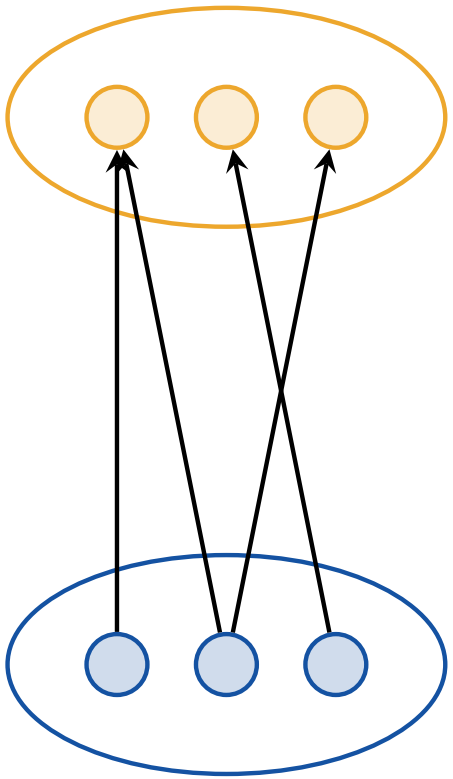


- $N$ number of translations
- $M$ number of sources

# One more problem



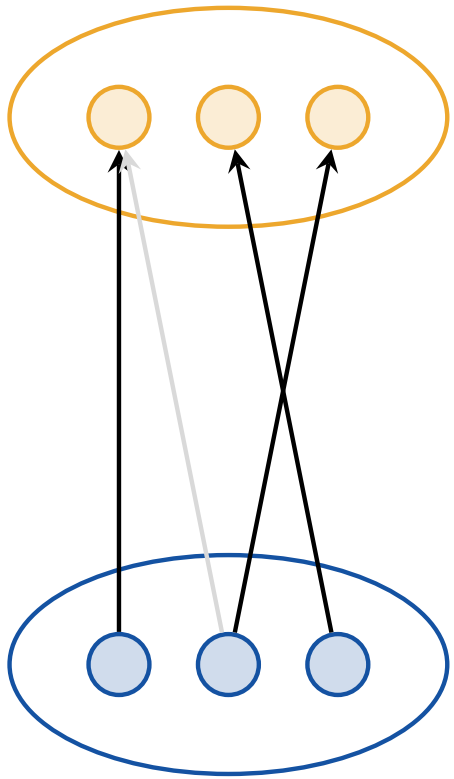
- $N$ number of translations
- $M$ number of sources



# One more problem



- $N$ number of translations
- $M$ number of sources



# Alignment



- Alignment can be used as a source of knowledge
- Alignment based features
  1. Aligned word list as lexical cues
  2. Frequency distribution of the most frequent aligned words

# More prospective features



1. Complexity of dependency structure
2. Entropy
3. Mutual Information
4. Number of different POS in a sentence (i.e.: number of adverbs)

# Outline



# Preliminary experiment

## Experiment setup

- Features
  1. Sentence length
  2. Type-token ratio
- classes
  1. Source
  2. Translation 1
  3. Translation 2
- Data: Europarl traning corpora
- Segment size: 500 sentences
- SVM setup parameter: Default in Weka



# Preliminary experiment

## Evaluation (Weighted average)

1. Precision: 0.684
2. Recall: 0.68
3. F-Measure: 0.655

## Confusion matrix

| | Source | Translation 1 | Translation 2 |
|---|---|---|---|
| Source | 166 | 5 | 5 |
| Translation 1 | 22 | 131 | 23 |
| Translation 2 | 65 | 49 | 62 |

# Sum up



# Reference

1. (Baker 1996): Baker, M., *Corpus-based translation studies: The challenges that lie ahead*. Pages 175–186 of: Somers, H. (ed), Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager. Amsterdam: John Benjamins.
2. (Pastor et al. 2008): Corpas, G., Mitkov, R., Afzal, N., Pekar, V., *Translation: Do they exist? A corpus-based NLP study of convergence and simplification*, In Proceedings of the AMTA, Waikiki, Hawaii.
3. (Hansen 2003): Hansen, S., *The Nature of Translated Text*, Saarbrucken: Saarland University.



# Reference

4. (Mehler 2008): Mehler, A., *Structural Similarities of Complex Networks: Computational Model by Example of Wiki Graphs*, Applied Artificial Intelligence, 22(7/8)
5. (Botafogo et al. 1992): Botafogo, R. A., Rivlin, E., and Shneiderman, *Structural analysis of hypertexts: Identifying hierarchies and useful metrics*, ACM Transactions on Information Systems, 10(2):142–180.
6. (Bernardini 2007): Bernardini, s., *Collocations in Translated Language. Combining parallel, comparable and reference corpora*, The Fourth Corpus Linguistics Conference, Birmingham, 27-30 July.



# Reference

7. (Smith and Senter 1967): Smith, E.A. and Senter, R.J *Automated Readability Index*, AMRL-TR,66-22, Wrright-Patterson AFB, OH:Aerospace Medical Division.
8. (Coleman and Liau 1975): Coleman, M. and Liau, T.L, *A computer readability formula designed for machine scoring*, Journal of Applied Psychology, Vol. 60, 283-284.
9. (Mehler et al. 2010): Mehler, A., Gleim, R. , Waltinger, U. and Diewald.N., *Time Series of Linguistic Networks by Example of the Patrologia Latina*, Proceedings of INFORMATIK 2010: Service Science, September 27 – October 01, 2010, Leipzig.