String Similarity Measures for Template Extraction

Natalia Elita

University of Hamburg

NATS Oberseminar, 07.06.07

Outline

- **Motivation**
- Similarity Matrix
  - String Similarity Measures
  - Indexing
- Template Extraction
- Conclusion
- Further work

Motivation -1-

(En) 1. The prosecution had charged Priebke with multiple and particularly ferocious homicide .

(De) 1. Die Staatsanwaltschaft hatte Priebke des mehrfachen , besonders grausamen Mordes beschuldigt .

(En) 2. In the course of the trial , lasting three months , Priebke had admitted to have shot to death two people himself .

(De) 2. Priebke hatte in dem 3 Monate dauernden Prozess zugegeben , 2 Menschen eigenhaendig erschossen zu haben .

Motivation -2-

(En) 1. **The** prosecution had charged **Priebke** with multiple and particularly ferocious homicide .
(En) 2. In **the** course of the trial , lasting three months , **Priebke** had admitted to have shot to death two people himself .

(De) 1. Die Staatsanwaltschaft ***hatte Priebke*** des mehrfachen , besonders grausamen Mordes beschuldigt
(De) 2. ***Priebke hatte*** in dem 3 Monate dauernden Prozess zugegeben , 2 Menschen eigenhaendig erschossen zu haben .

Motivation -3-

(De) 1. **Weitere Informationen finden Sie unter Sicherheitseinstellungen auf Seite NUM .**
(De) 2. **Weitere Informationen** hierzu **finden Sie unter Sicherheitseinstellungen auf Seite NUM .**

(En) 1. ***For further information , see Security settings on page NUM***
(En) 2. ***For further information , see Security settings on page NUM .***

Outline

- Motivation
- **Similarity Matrix**
  - String Similarity Measures
  - Indexing
- Template Extraction
- Conclusion
- Further work

Similarity Matrix

- Similarity Matrix:
  - For a monolingual corpus with *N* sentences, the Similarity Matrix *s* is formally defined:

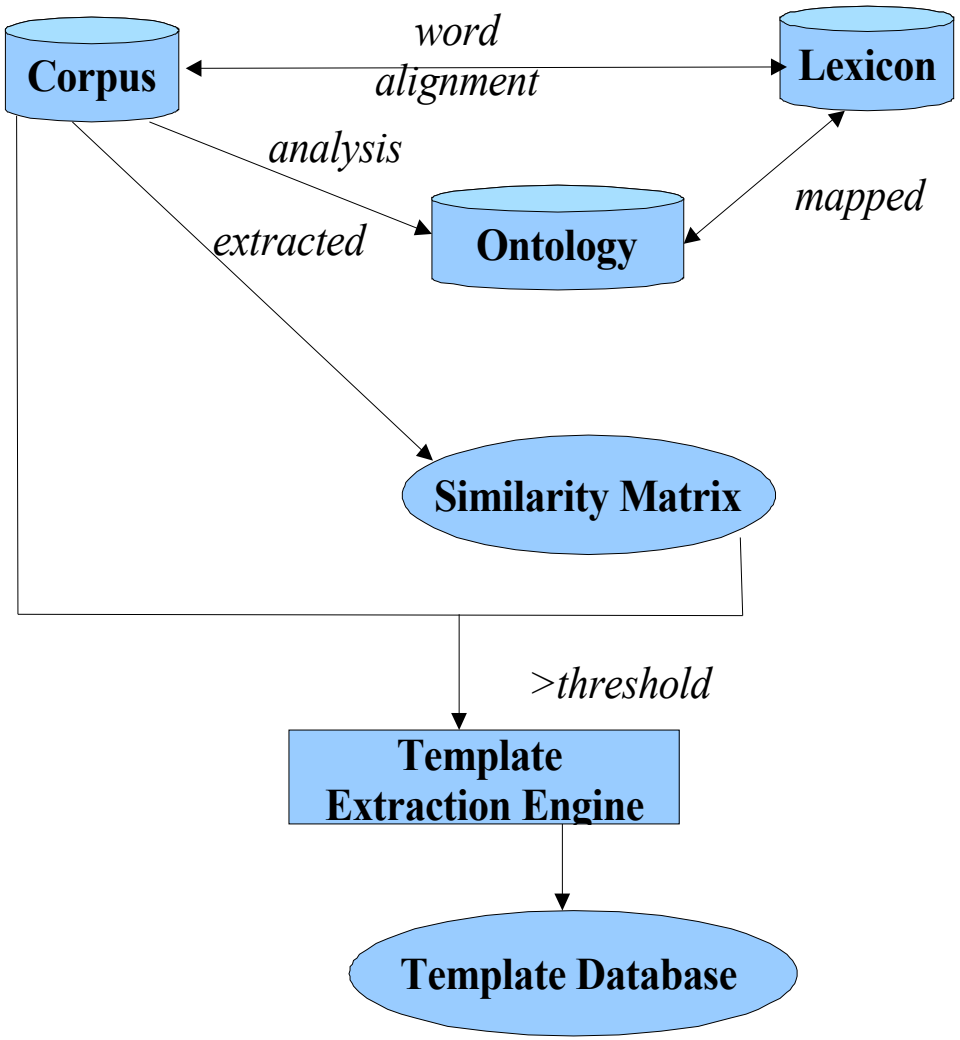
s(i,j)=0, for j<i, 1<=i,j<=N;

s(i,i)=1, for 1<=i=N;

s(i,j)=BSM(sentence$_i$, sentence$_j$), for j >i, 1<=i, j<=N, where *BSM* = Best Similarity Measure

  - to reduce the search space
  - to find candidates for templates
  - to observe the need of semantics
- Indexing
  - to reduce the search space

Similarity Matrix



	s1	*s2*	*s3*	...	*sn*
s1	1	0.85	0.02	...	0.15
s2	0	1	0.12	...	0.96
s3	0	0	1	...	0.48
...	0	0	0	1	0.50
sn	0	0	0	0	1

Template

- generalization of sentences that are translations of each other, where sequences of one or more words are replaced by variables, with alignments between the resulting word sequences and/or variables made explicit
  - E.g (SL)**Tfa** V_i **Tfb** V_{i+1} **Tfc** <---> (TL) Vi **Tfd** V_{i+1} **Tfe ,**

where **Tfx – text fragment x**

V_i – *variable i*

Problem description

- Given a sentence aligned corpus, find sentences that are similar enough to become candidates for translation templates
  - no syntactic annotation of the corpus
  - no other linguistic resource
  - similarity on the surface form only

Outline

- Motivation
- Similarity Matrix
  - **String Similarity Measures**
  - Indexing
- Template Extraction
- Conclusion
- Further work

String Similarity Measures

- String Similarity measures are used in applications:
  - Spell check
  - Text prediction
  - Translation Memories
  - EBMT (matching)
  - ...

Types

- character-based
  - similarity at the character level
- token-based
  - similarity at the token level
- hybrid
  - token based similarity first applied, then character based on each similar token

String Similarity Measures under consideration

- 20 string similarity measures
  - 18 – SymMetrics package*
    - 10 character based, 5 token based, 3 hybrid
- 2 – new
  - token-based
    - Common Words (CW)
    - Adapted Levenshtein Distance (ALD)

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

New Token based Measures -1-

- Common Words (CW):
  - number of common tokens for two given strings *s1 and s2*

e.g:
(s1) Writing ***and*** *sending* ***a multimedia message***
(s2) Reading ***and*** *replying to* ***a multimedia message***

CW = 4 [and a multimedia message]

New Token based Measures - 2-

- *Adapted Levenshtein Distance (ALD)*
  - *For the given two strings s1 and s2:*
    - *Token Levenshtein Distance (TLD) is the traditional Levenshtein Distance, but on token level;*
    - *The maximal number of tokens of s1 and s2 is determined;*
    - *The obtained value is normalized to get values between 0 and 1.*

ALD (example)

$$ALD(s_1,s_2)=1-\frac{TLD}{2*\max(Length(s_1),Length(s_2))}$$

(s1) Writing ***and*** *sending* ***a multimedia message***

(s2) Reading ***and*** *replying to* ***a multimedia message***

TLD = 3

max(length(s1),length(s2)) = 7

ALD = 1-(3/14)= 0.78

Thresholds

- experimentally established
  - identical strings (1)
  - completely different strings (0)
  - substrings
    - word order
    - length of strings

Thresholds: Character-based

- TagLink Token = 0.5
- Euclidean Distance = 0.5
- Smith-Waterman = 0.6
- Smith-Waterman-Gatoh = 0.6
- Jaro = 0.7
- Jaro Winkler=0.7
- Needlemann-Wunch= 0.7
- Levenshtein Distance = 0.75
- Dice Similarity=0.75
- Cosine Similarity= 0.75

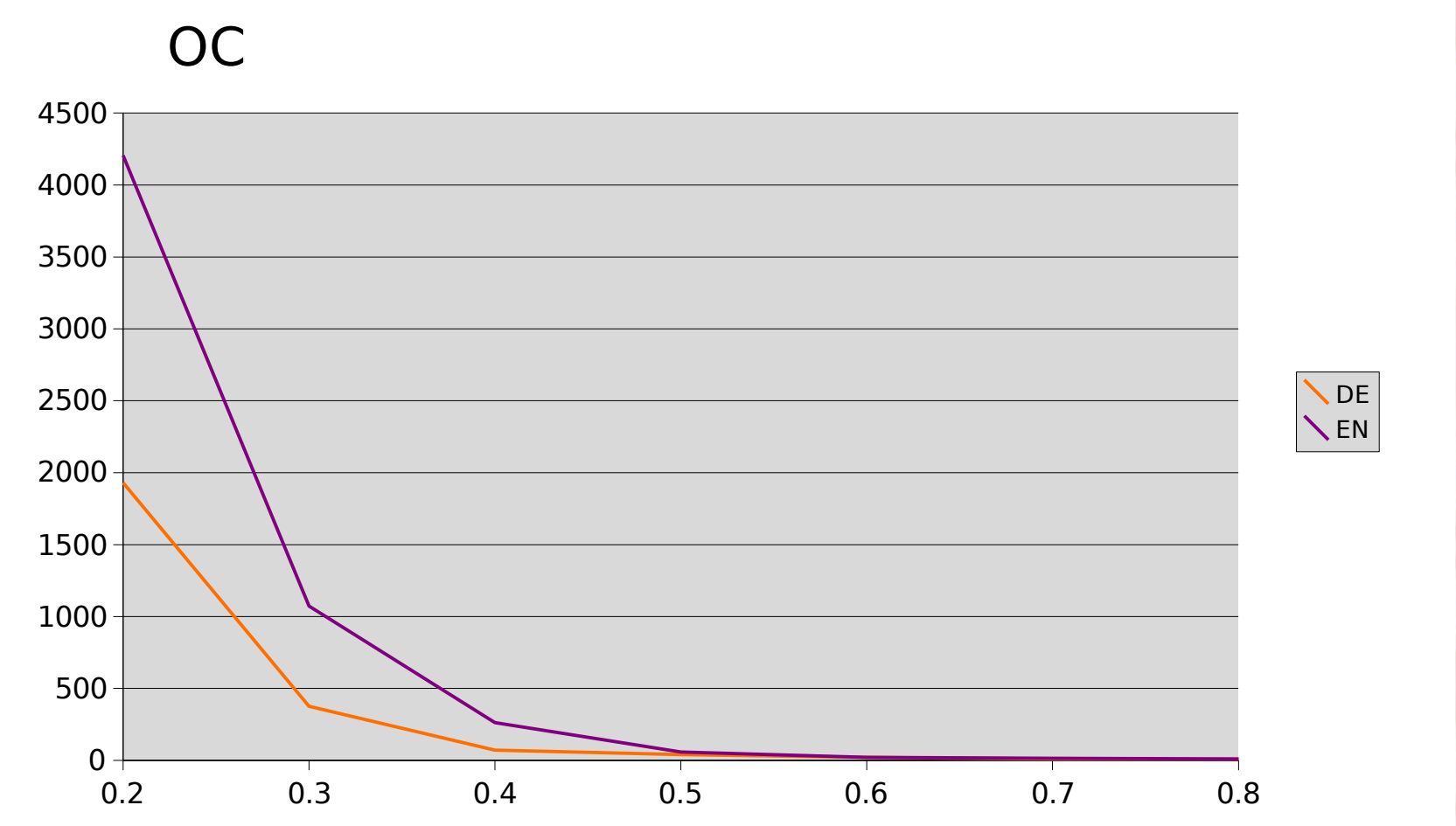
Thresholds: Token-based

- Common Words (CW) = 5
- Adapted Levenshtein Distance = 0.7
- Matching Coefficient = 0.55
- Block Distance = 0.6
- Jaccard Similarity = 0.45
- Overlap Coefficient (OC) = 0.66
- Q-Grams Distance = 0.65

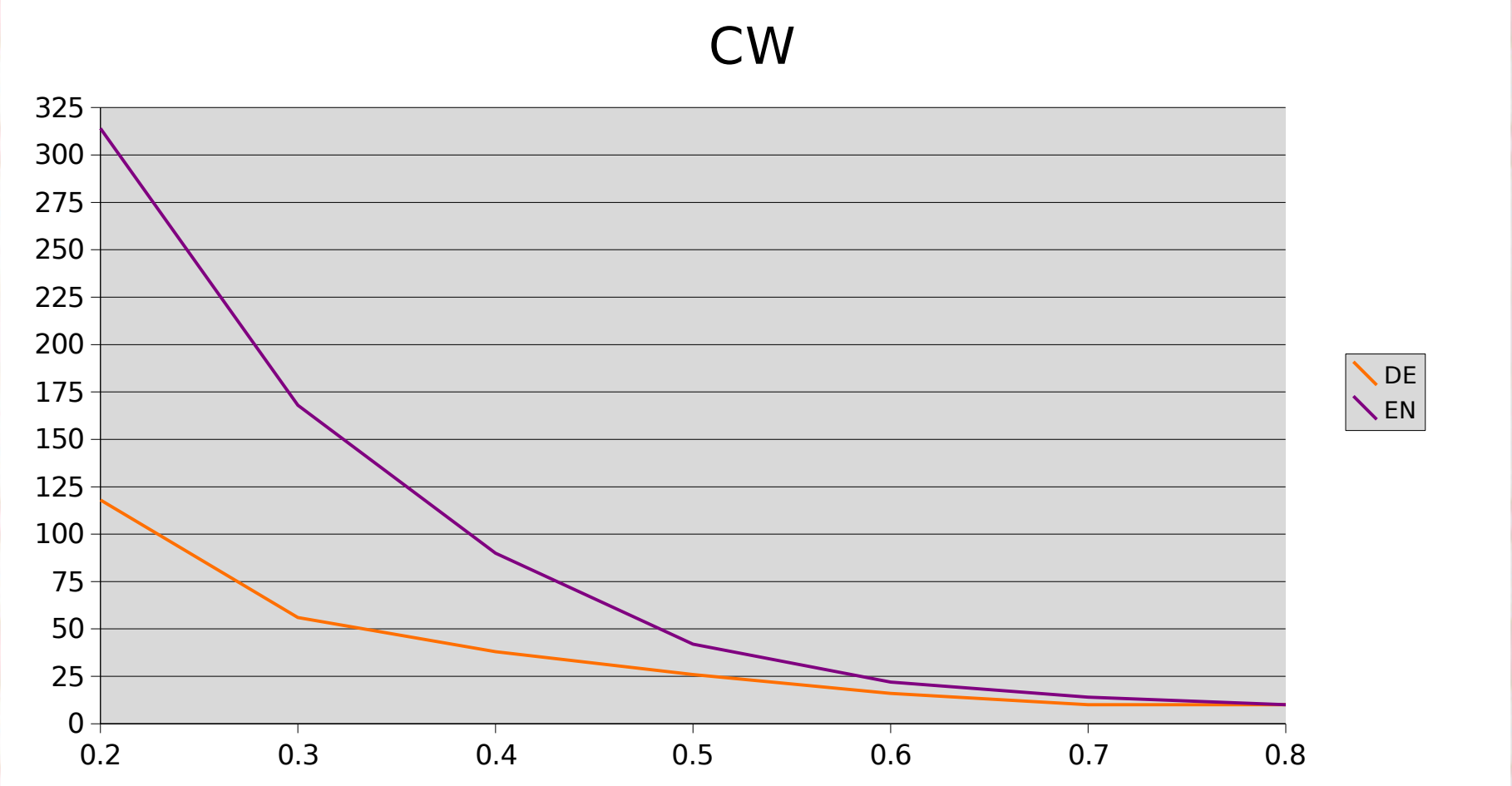
Thresholds: Hybrid

- Monge-Elkan = 0.9
- Chapman Ordered Name Compound Similarity = 0.75
- TagLink = 0.7

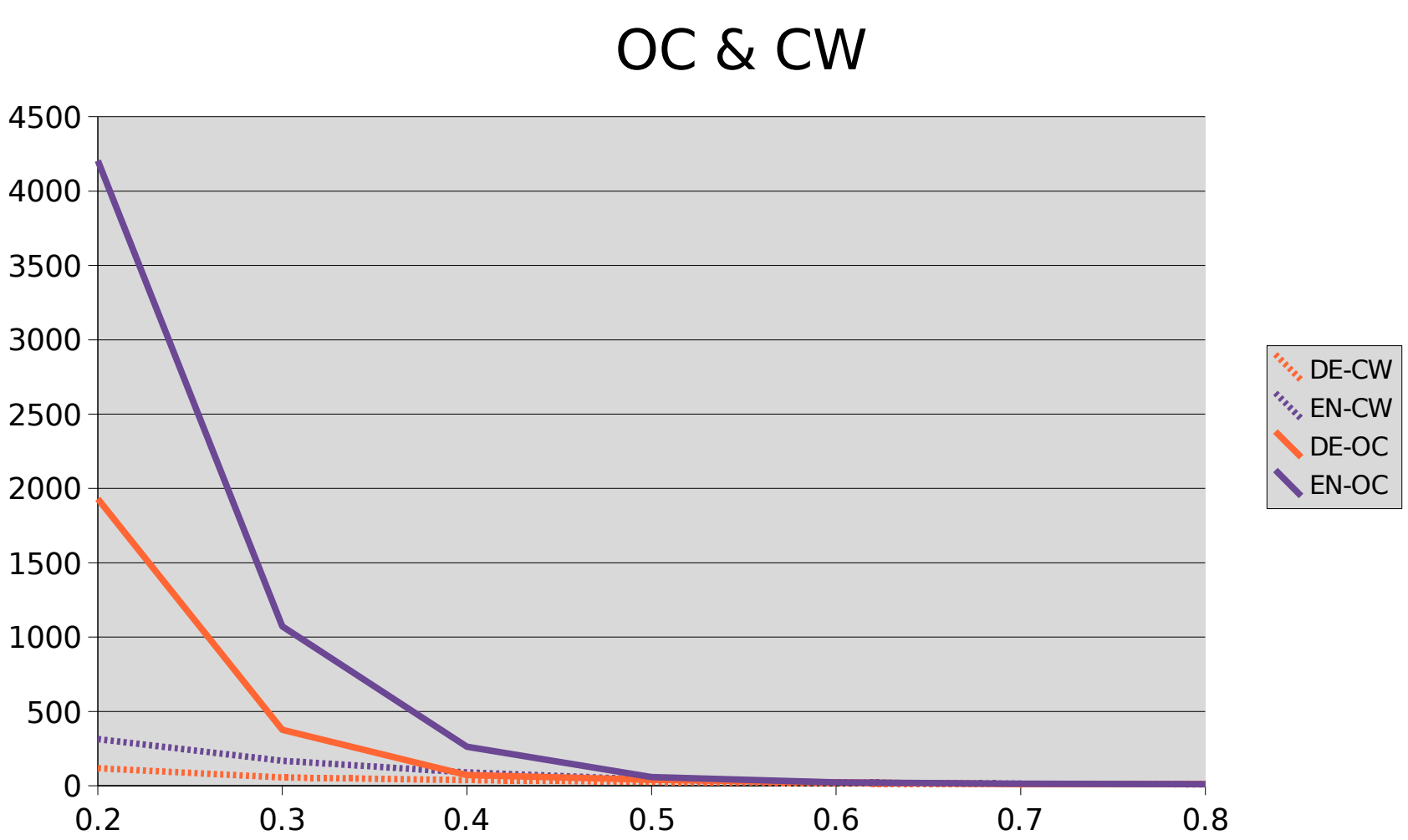
Thresholds/Candidates for Templates (OC)



Thresholds/Candidates for Templates (CW)



Thresholds/Candidates for Templates



Experimental settings

- corpus: technical
- languages: De, En, Ro
- 100 sentences
  - to make observations, assumptions
  - manual evaluation

Experiments -1-

Token-based	**Ge**	**En**	**Ro**
CW	4	11	11
Matching coefficient	12	10	9
Block Distance	13	12	13
Jaccard Similarity	12	10	9
OC	**24**	**19**	**25**
Q-Grams Distance	9	9	6
Total	74	71	73
Unique pairs	**26**	**30**	**31**

Experiments -2-

Character-based	**Ge**	**En**	**Ro**
Levenshtein Distance	1	3	2
Dice Similarity	5	4	3
Cosine Similarity	5	4	3
Euclidean Distance	5	4	3
Jaro	35	32	56
Jaro-Winkler	86	72	109
Needleman-Wunch	24	40	22
SW	83	82	49
SW-Gotoh	**107**	**103**	**73**
Tag Link Token	70	67	62
Total	421	411	382

Experiments -3-

Hybrid	**Ge**	**En**	**Ro**
CONC	**48**	**48**	**29**
Tag Link	19	17	19
Total	67	65	48
Unique pairs	**58**	**59**	**40**

Observations

- Character-based measures too slow and depend very much on the length of the strings to be compared
  - e.g. 300 sentences (De,Ro) ~ 7 minutes
- Hybrid methods – perform not so well in case of German compound nouns
- Token-based – the most useful for the template extraction
  - Common Words and Overlap Coefficient

Observations -2-

- Common Words – the number of common tokens two strings have
  - no word order is taken into account
- Overlap Coefficient (OC) – the metric which determines to what degree is one string a substring of another:

$$OC(s_1, s_2) = \frac{(|s_1 \wedge s_2|)}{\min(|s_1|, |s_2|)}$$

where: |s| - number of tokens in s,

$|s_1 \wedge s_2|$ - number of common tokens in s_1 and s_2

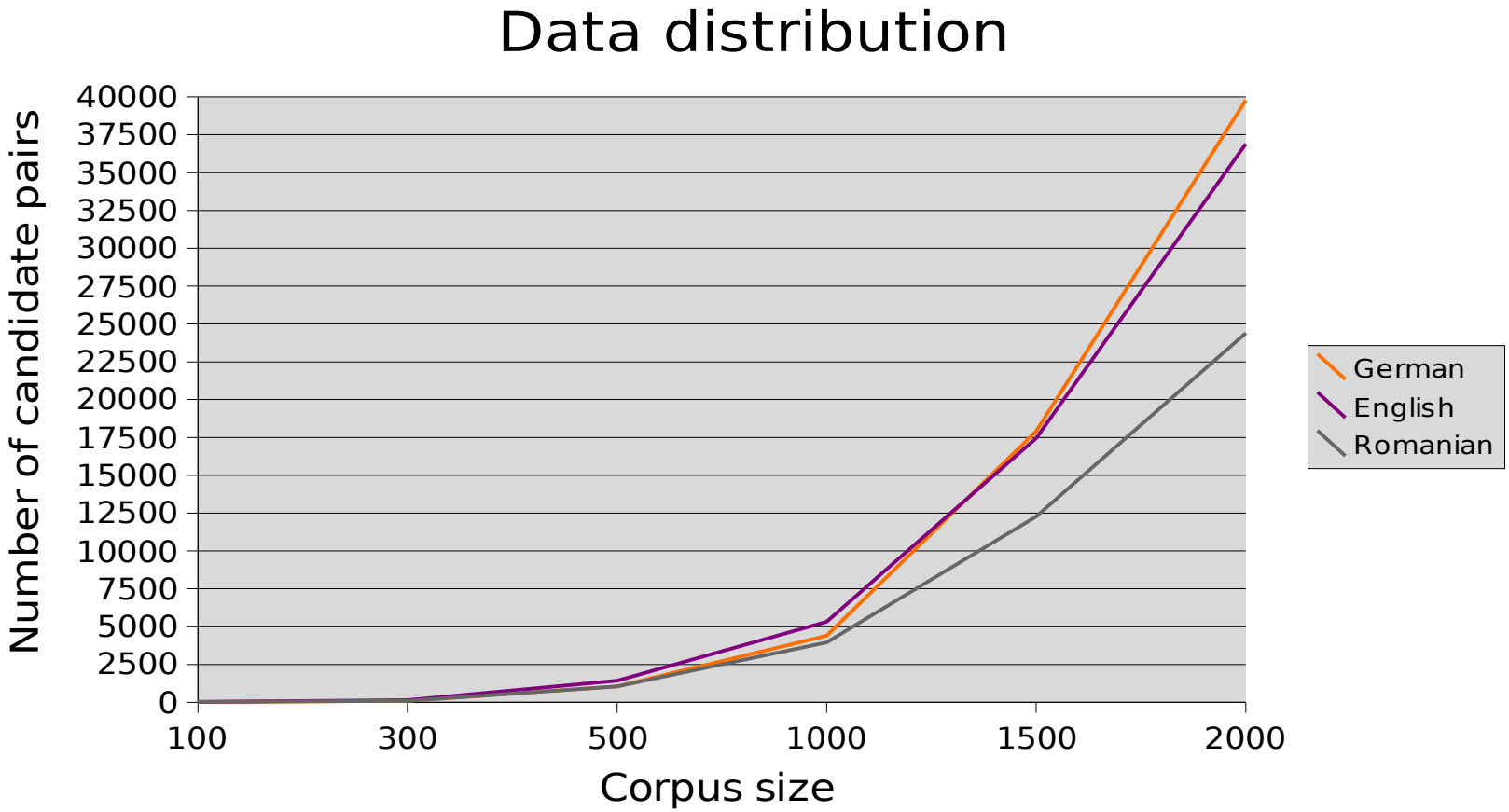
Observations

- CW + OC used to build the Similarity Matrix
  - Thresholds: CW =3; OC = 0.5;
  - Experiments made on sets
    - in different languages
    - of different size
    - of different corpus type

Experiments

- goal: for each language, see how the number of similar sentences changes with the size of the corpus
  - corpus type: technical
  - corpus size: up to 2000 sentences
  - languages: De, En, Ro

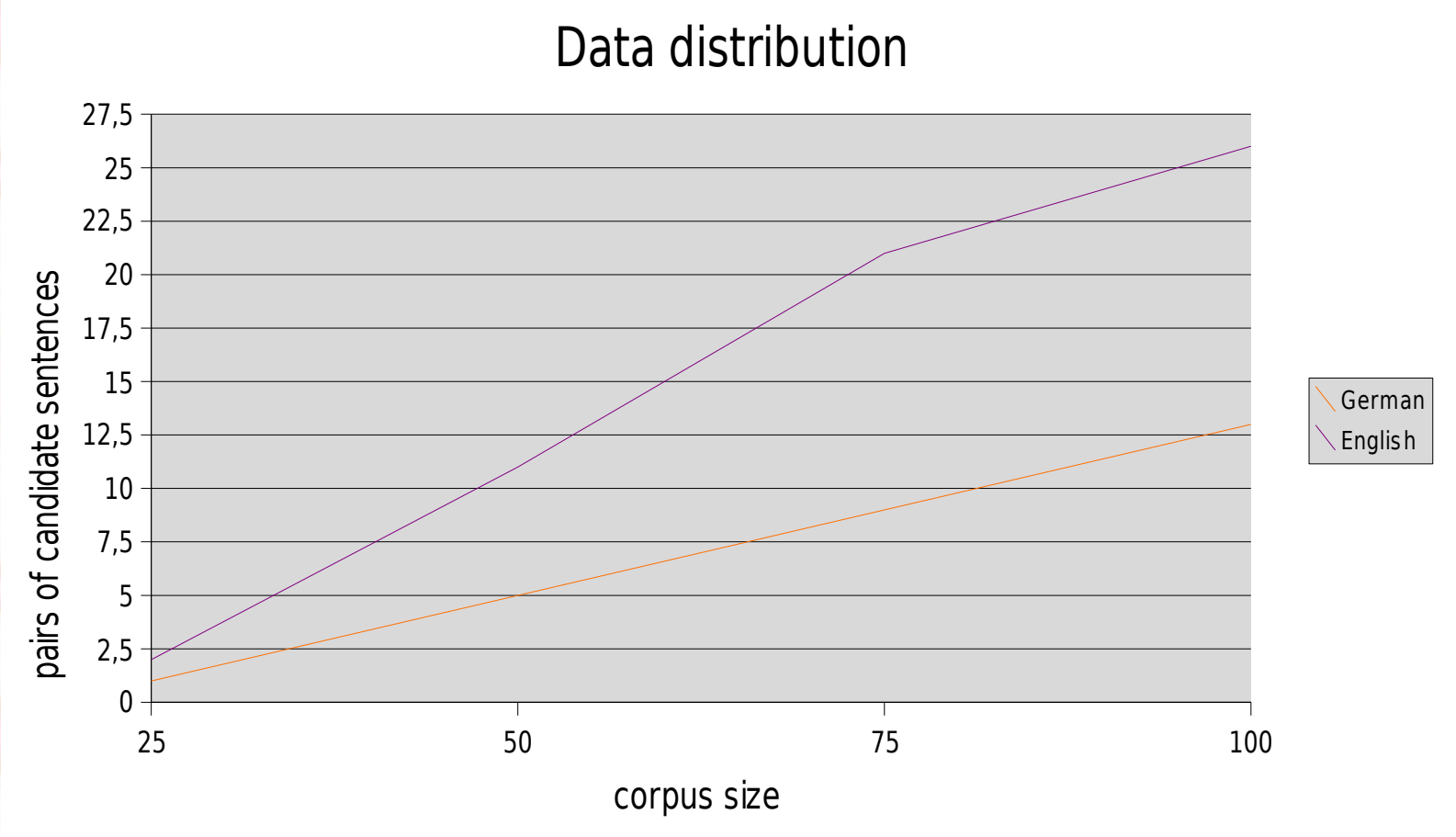
Experiment -1-



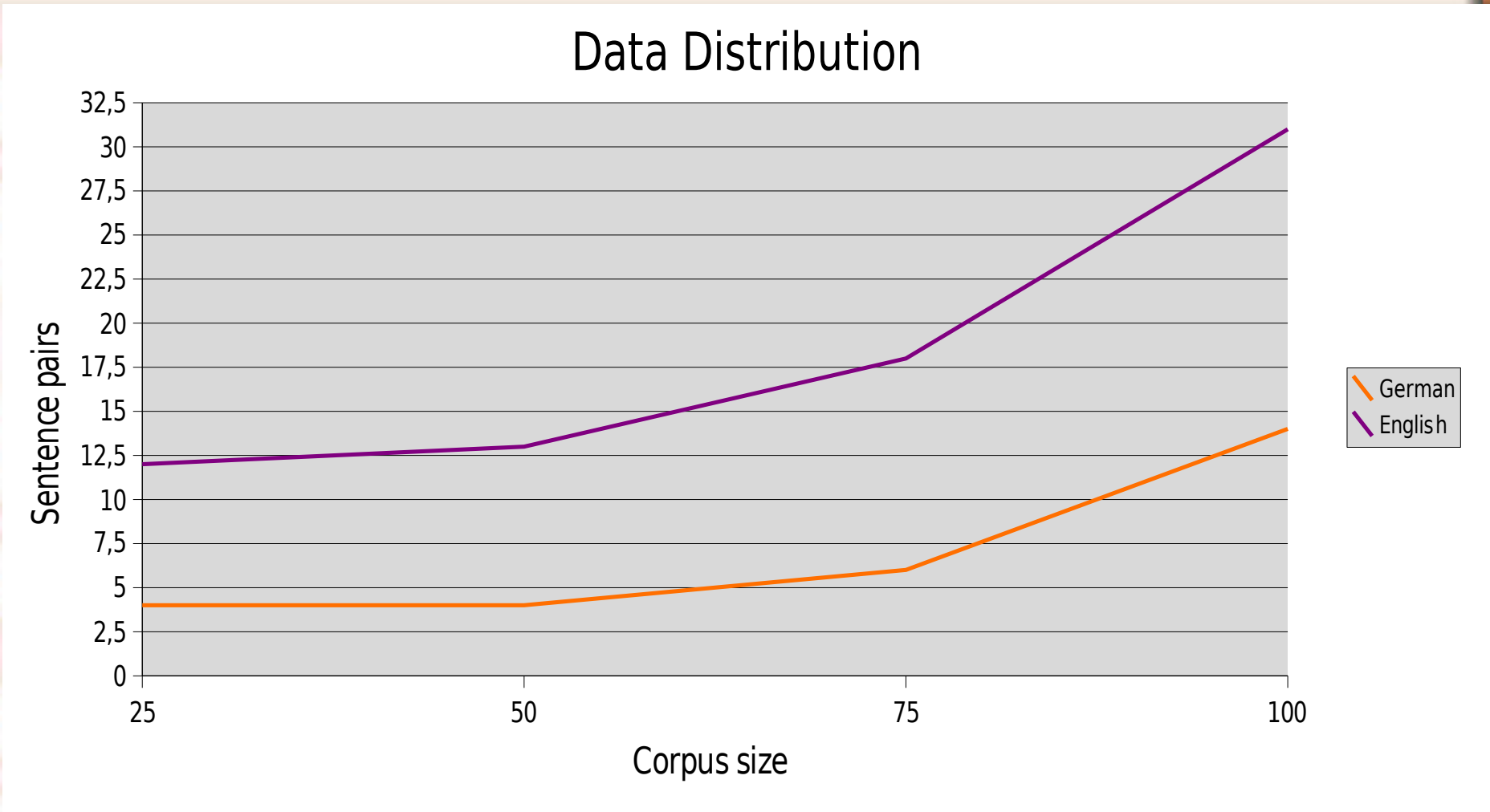
Experiment -2-

- corpus dependency
  - up to 100 sentences
    - news and technical corpora
    - languages: De, En

Experiment -2-(News)



Experiment -2- (Technical)



Outline

- Motivation
- Similarity Matrix
  - String Similarity Measures
  - **Indexing**
- Template Extraction
- Conclusion
- Further work

Index vs Similarity Matrix

- Search for similar sentences
  - *n*(n-1)/2* comparisons have to be made, where *n* is the number of sentences in a corpus
    - e.g: corpus of 100 sentences – 4950 comparisons
- Index

Corpus type	Language	Corpus size	Search space
News	En	100	2001
News	De	100	1390
Technical	En	100	479
Technical	De	100	456

Outline

- Motivation
- Similarity Matrix
  - String Similarity Measures
  - Indexing
- **Template Extraction**
- Conclusion
- Further work

Baseline System

Language neutral recursive machine learning algorithm based on principle of similar distributions of strings:

Source Language and Target Language strings that co-occur in two (or more) sentence pairs of a bilingual corpus are likely to be translations of each other

Problems

Proved to have serious limitations:

-the templates obtained are often not translations;

-no template is learned if different lexical items are used - semantics would be extremely useful in this case;

-big memory problems for a small corpus of 400 sentences;

-useful information is lost.

Example -1-

Given 2 sentences in English:

12: **The** discussion around **the** envisaged major **tax reform** continues .
16: **The** head of the FDP parliamentary group , Mr. Solms , however , has deviated from the FDP 's demand to enact **the tax reform** as early as 1998 .

The sequence of common elements: **[the the tax reform]**

Example -1-

Generalized template fragments of these 2 sentences:
[**The** *V1* **the** *V2* **tax reform** *V3*] (12)
[**The** *V4* **the tax reform** *V5*] (16)
Where:
V1 = "discussion around"
V2 = "envisaged major"
V3 = "continues"
V4 = "head of the FDP parliamentary group , Mr. Solms , however , has deviated from the FDP 's demand to enact"
V5 = "as early as 1998"

Example -1-

The translations into German:

12: Die Diskussion um **die** vorgesehene grosse **Steuerreform** dauert an.

16: Der FDP - Fraktionsvorsitzende im Bundestag , Solms , ist von der Forderung der Liberalen abgerueckt , **die Steuerreform** schon 1998 in Kraft zu setzen.

The sequence of common elements: [**die Steuerreform**]

The sequence contains only 2 elements --> threshold established at 3

Solution?

For a given SL corpus:

1. Index created for each sentence in the corpus.
2. Similarity matrix build for the corpus:

pairs of "similar" sentences with the sequence of common elements greater or equal to three are extracted;

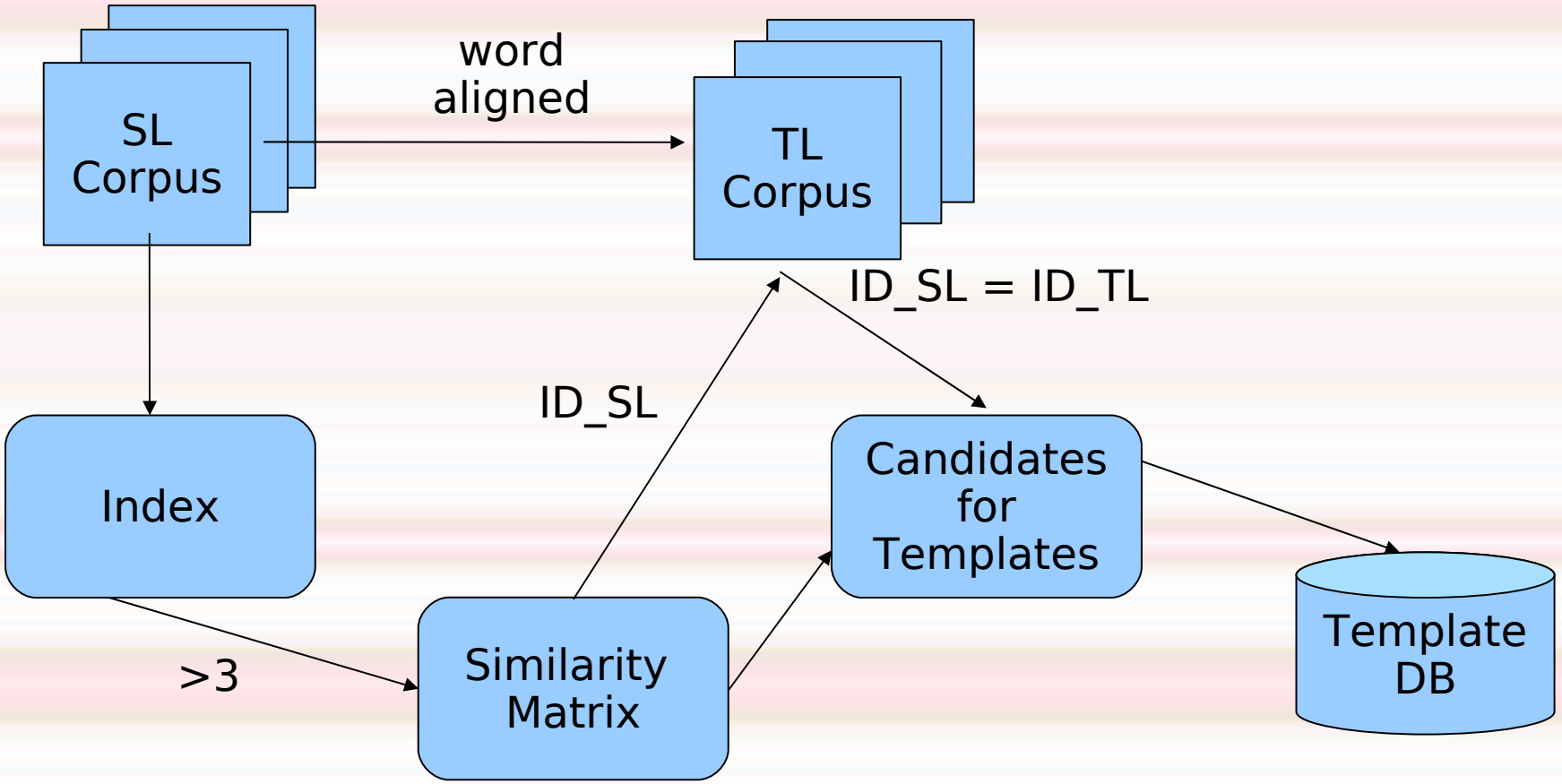
for each pair of similar sentences: their TL (by the sentence ID) counterparts are retrieved

for each pair of sentences SL and TL with the same IDs are (word) aligned;

Corresponding TL counterparts of the sequence of common elements are found;

SL and TL parts combined into a template

Algorithm


SL
Corpus
word
aligned
TL
Corpus
ID_SL = ID_TL
ID_SL
Index
Candidates
for
Templates
Template
DB
>3
Similarity
Matrix

Another example

For the given sentences in SL (English):

26: Wage **conflict in *retail*** *business grows*
27: The **conflict in** the wage negotiations in the ***retail*** *industry has extended* to North Rhine Westphalia .

and the translations into TL (Cerman):

26: **Tarifkonflikt im *Einzelhandel*** *weitet sich aus*
27: Der **Tarifkonflikt im *Einzelhandel*** *hat sich* auf Nordrhein - Westfalen *ausgeweitet* .

Variables - 1 -

8. [**wage conflict in retail** *V1*]--> [**Tarifkonflikt im Einzelhandel** *V11* **sich** *V21*]

V1 = "business grows"
V11 = "weitet "
V21 = "aus"

Variables - 2 -

9. [*V1* **conflict in** *V2* **wage** *V3* **retail** *V4*] -->[*V11* **Tarifkonflikt im Einzelhandel** *V21* **sich** *V31*]

V1 = "The"
V2 = "the"
V3 = "negotiations in the"
V4 = "industry has extended to North Rhine Westphalia"

V11 = "Der"
V21 = "hat"
V31 = "auf Nordrhein - Westfalen ausgeweitet"

Alignment

8. [wage conflict in retail *V1*]--> [Tarifkonflikt im Einzelhandel V11 sich *V21*]

V1 = "business grows"

V11 = "weitet "
V21 = "aus"

Problems to solve

- tense/aspect:
  - grows vs has extended
- semantics:
  - retail business vs retail industry
  - grows vs extends

Solution to semantics: WordNet -1-

retail business vs retail industry
WordNet:

Industry is a direct hyponym of **business** as seen from the WordNet:

S: (n) commercial enterprise, business enterprise, **business** (the activity of providing goods and services involving financial and commercial and industrial aspects) "computers are now widely used in business"

* **direct hyponym** / full hyponym

o S: (n) **industry**, manufacture (the organized action of making of goods and services for sale) "American industry is making increased use of computers to control production"

Solution to semantics: WordNet -2-

grow/extend - no direct connection found;
indirectly - grow -->expand (direct troponym); extend -->expand (verb group);

Problem:

How do I know I chose the right sense of business?
Difficult even for a human to decide which synset is appropriate.

Solution to semantics: FrameNet

FrameNet:

Industry is the lexical unit (LE) belonging to the frame **Fields**, and LE **Business** belongs to the **Business frame**.

Grow/Extend:

LE	Frame
(1) grow.v	Expansion
(2) grow.v	Cause_expansion
(3) grow.v	Becoming
(4) grow.v	Change_position_on_a_scale

Solution to semantics: FrameNet

LE **extend** contained in the frame **Change_event_duration.**

Definition: In this frame, an Agent or Cause changes the duration of an Event. The Event will then take place for a New_duration, rather than the Initial_duration. This can be done with by certain Means, in a certain Manner or to a certain Degree.

In my opinion, in our context - the meaning of "extend" does not correspond to the definition of the frame, as certainly an idea of space is expressed by it.

Another example (need of semantics) -1-

Given the two pairs of sentences:

26: **Wage** *conflict* **in** ***retail*** *business* grows
97: **Wage** *dispute* **in** ***retail*** *sector*

26: Der **Tarifkonflikt** *im Einzelhandel* hat sich auf Nordrhein - Westfalen ausgeweitet .
97: **Tarifkonflikt** *des Einzelhandels*

Another example (need of semantics) -2-

- ***WordNet:***

conflict/dispute *- the same synset in WordNet:*

S: (n) ***dispute,*** *difference, difference of opinion,* ***conflict*** *(a disagreement or argument about something important) "he had a dispute with his wife"; "there were irreconcilable differences"; "the familiar conflict between Republicans and Democrats"*

business/ business sector *- the same synset in WordNet*

Another example (need of semantics) - 3-

- ***FrameNet***

*LE **dispute** - in Quarrelling frame;*
*LE **conflict** - in Hostile Encounter frame;*

LE business - in Business frame
LE sector - in Fields frame

Evaluation -1-

Experiments done with the news corpus (100 sentences)

A total of 53 template fragments were extracted, only 16 of them can be combined in a full template - by the sentence IDs the fragments were extracted from.

Semantics:
Noticed to be useful in 8 template fragments

Evaluation - 2 -

Errors:

- Extracted fragments not translations - 4 cases
- No fragments learned because of:
  - Common Words Threshold (De) – 15 cases
  - Overlap Coefficient Threshold (En) – 5 cases
  - Spelling errors – 1 case
  - Paraphrase – 2 cases

Outline

- Motivation
- Similarity Matrix
  - String Similarity Measures
  - Indexing
- Template Extraction
- **Conclusion**
- Further work

Conclusion

- Similarity matrix used to find candidates for templates
  - Common Words and Overlap Coefficient as similarity criteria
  - Index used to reduce the search space
- Generalization of similar sentences into translation templates needs semantic information

Outline

- Motivation
- Similarity Matrix
  - String Similarity Measures
  - Indexing
- Template Extraction
- Conclusion
- **Further work**

Further work

- **Decisions on templates**:
  - Generalize on at least two sentences?
  - If common tokens are in different order, on which sentence should the generalization be made?
  - Variables: one token per variable?
- **Extract templates without semantics**
  - Decide on the source of semantics
- **Add semantic information**
- **Extract templates with semantics**

Thank you!

Questions? Suggestions?