Cross-language High Similarity Search using a Conceptual Thesaurus

Parth Gupta[1], Alberto Barrón-Cedeño[2] and Paolo Rosso[1]

[1]Universitat Politècnica de València, Spain
[2] Universitat Politécnica de Catalunya, Spain

September 19, 2012

Outline

Introduction

- The task of cross-language high similarity search refers to the identification of documents that are duplicates or share very similar information in two different languages.
- Some examples
  - Wikipedia articles in multiple languages
  - news stories in different languages covering the same event
  - cross-language cases of plagiarism
  - translated documents etc.
- In the literature, also referred as
  - Cross-language pairwise similarity search
  - Cross-language mate retrieval
  - Cross-language near duplicate search

Conceptual Thesaurus (Domain specific)

- Has often a multi-word structure
- Tries to exhaustively cover omnipresent concepts of the domain
- Eurovoc[1]
  - Emerged from European Parliamentary proceedings
  - Contains 6,797 multilingual concepts in 22 languages
  - Span across 21 domains of European Parliament activities

[1] http://eurovoc.europa.eu/

Eurovoc

English	Spanish	German
action for failure to fulfil an obligation	recurso por incumplimiento	Klage wegen Vertragsverletzung
extra-community trade	intercambio extracomunitario	außergemeinschaftlicher Handel
sexual harassment	acoso sexual	sexuelle Belästigung

Eurovoc

- Domain of concepts
  - Politics
  - Intenational relations
  - European community
  - Law
  - Economics
  - So on..

Assigning these concepts to Wikipedia documents or Shakespeare stories?

Method - Cross-language Conceptual Thesaurus based Similarity (CL-CTS)

- Represent documents as a vector of concepts
- Concept assignment is the least trivial part
- Challenge: Exploit a domain specific CT for all the corpora
- Assignment of concepts according to their verbatim occurrence in the document gives very bad results [Pouliquen et al.2006]
- Assign a concept to a document if it "triggers the concept"

Method contd.

- Heuristic: *The terms together are highly domain dependent but alone are domain independent.*
- For example, "community" and "trade" compared to "community trade"

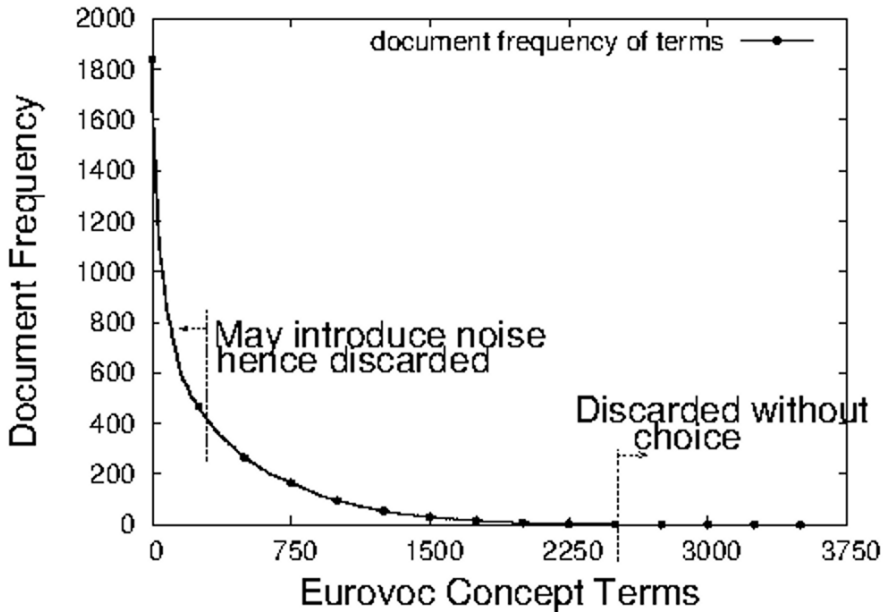
Concept Assignment

- Sum of the term frequencies (TF) of the terms in the concept in the Doc
- Stopword removal + stemming
- Filter the terms based on the discriminative power in the corpora

Method contd.

- All the concepts do not help in similarity estimation - Hence Reduced Concepts (RC)

  - Reduces the comparison vocabulary drastically
  - Domain independent threshold $0 < df(t) < \beta$
  - Automatic domain adaptation (Football in "Sports" and "Society and Culture")



Method contd.

- Concern - The concepts are limited and are common across even slightly relevant documents
- To overcome the limitation of conceptual similarity estimation, we use Named Entities in similarity too
- n-gram similarity of NEs - simplest method
- NEs act as discriminative features - e.g. Wikipedia page of Rome vs. Madrid

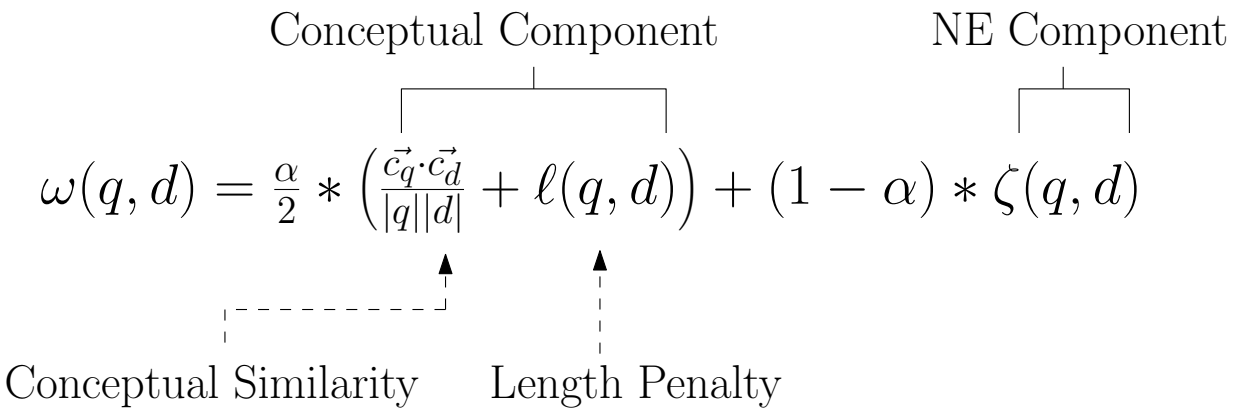
Method contd.

- Sometimes high similar documents are parallel and the task is to find the parallel document for the given document
- A pattern in length is noticed for parallel documents across languages [Pouliquen et al.2006]
- we use the same "length panelty"

$$\text{len}(\text{parallel}(d_q)) = f(\mu, \sigma, \text{len}(d_q))$$

Method contd.

- The similarity function



Compared with

1. Cross-language Alignment based Similarity Analysis (CL-ASA) [Barrón-Cedeño et al.2008, Pinto et al.2009]
2. Cross-language Character n-grams (CL-CNG) [Mcnamee and Mayfield2004]

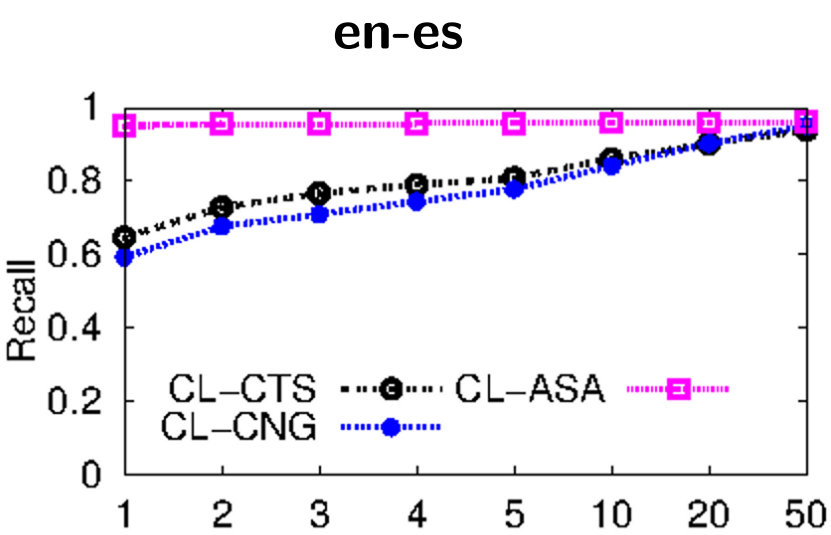
Datasets

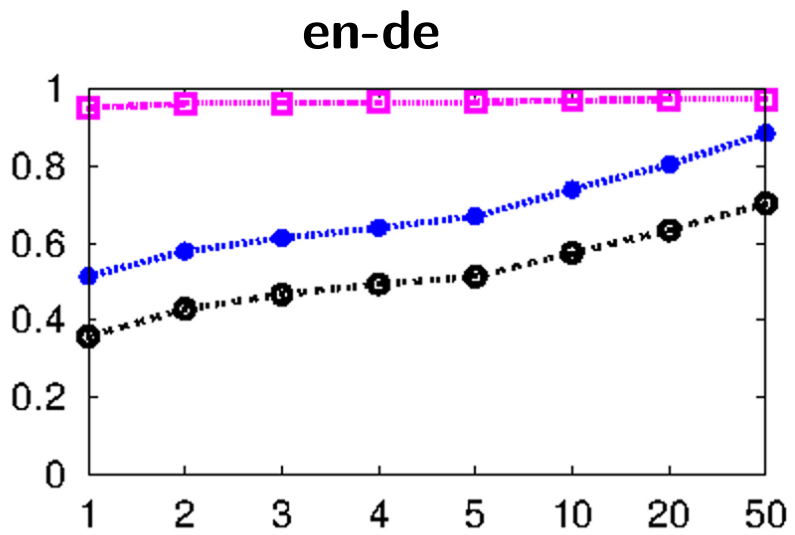
- JRC-Acquis (JRC)
  - Nature: related to European Commission activities
  - Size: 10,000 in each language
  - Type: Parallel
- PAN-PC-2011 (PAN)
  - Nature: Project Gutenberg (artificially created cross-language plagiarism cases)
  - Size: 2920 (en-es) and 2222 (en-de)
  - Type: Noisy parallel
- Wikipedia (Wiki)
  - Nature: General Wikipedia pages
  - Size: 10000 in each language
  - Type: Comparable

Datasets contd..

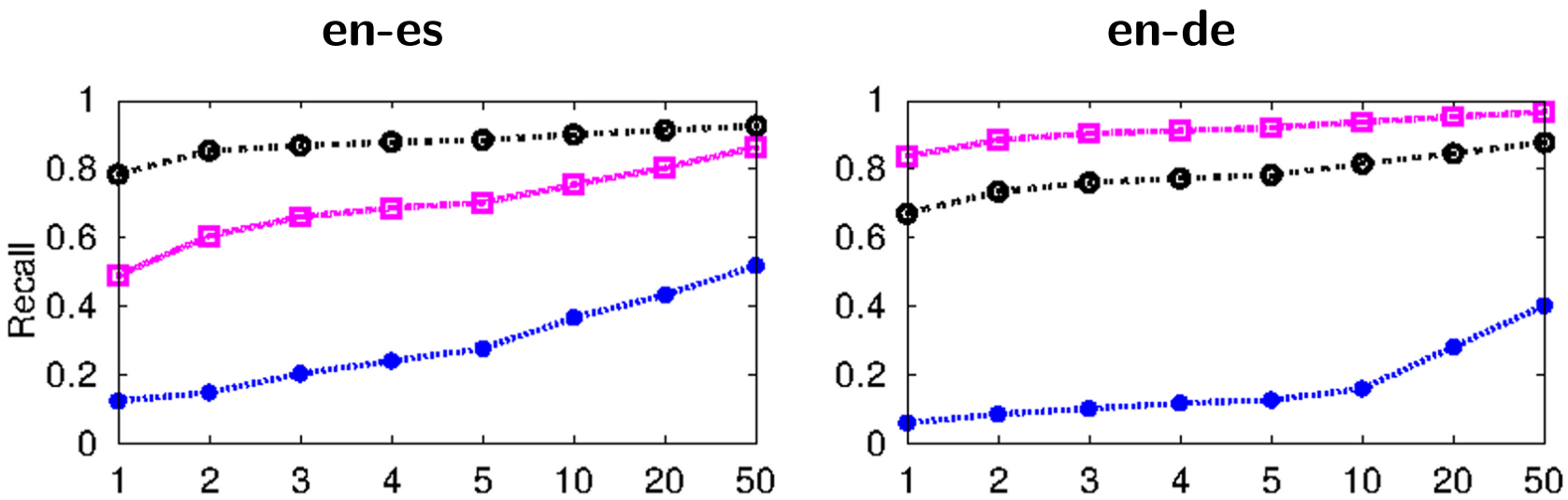
- Vocabulary shared by Eurovoc and JRC is higher than that of Eurovoc and PAN or Wiki.

Results : JRC

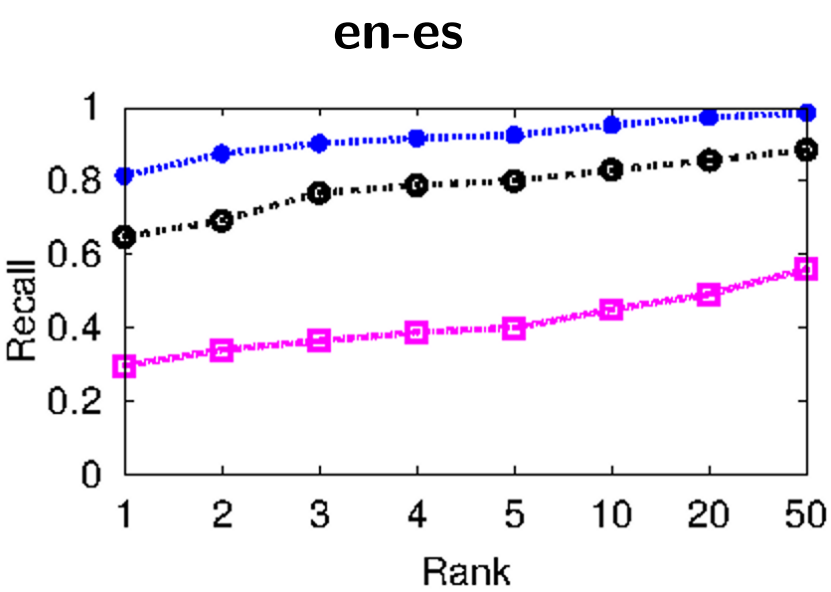

en-es
Recall
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
10
20
50
CL–CTS
CL–ASA
CL–CNG

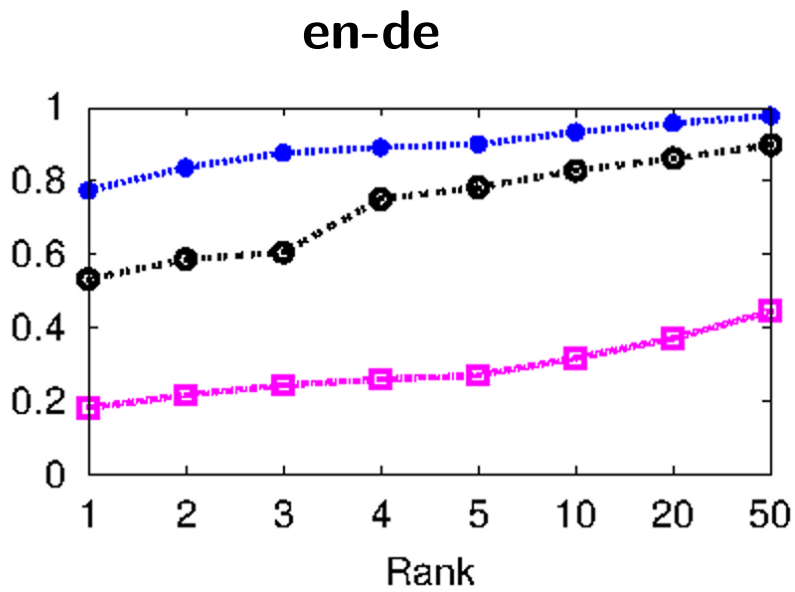

en-de
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
10
20
50

Results : PAN



Results : Wiki





Analysis

- Performance of CL-CTS with reduced concepts is much higher compared to inclusion of all concepts
  - R@1 0.02 $\rightarrow$ 0.58 (JRC en-es)
- Inclusion of NE component usually improves the performace except JRC - Interesting!
- CL-ASA and CL-CNG exhibit very corpus dependent performace.
- German stays more difficult compared to Spanish (compounding of the words needs better care)

Analysis: Further characterizing the corpora

- JRC
  - Parallel corpus
  - high amount of NEs
  - NEs are mostly of type ORG and LOC which are appear quite identically in many documents
- PAN
  - Cross-language plagiarism cases artificially generated using SMT and/or manual correction - Noisy Parallel
  - documents are related to literature - contains far more natural language terms compared to NEs
  - NEs are mostly of type PERSON and is much diverse across documents

Analysis contd.

- Wiki
  - Generic documents - comparable
  - Lots of NEs, but diverse

- We investigated the distribution of NEs among corpora

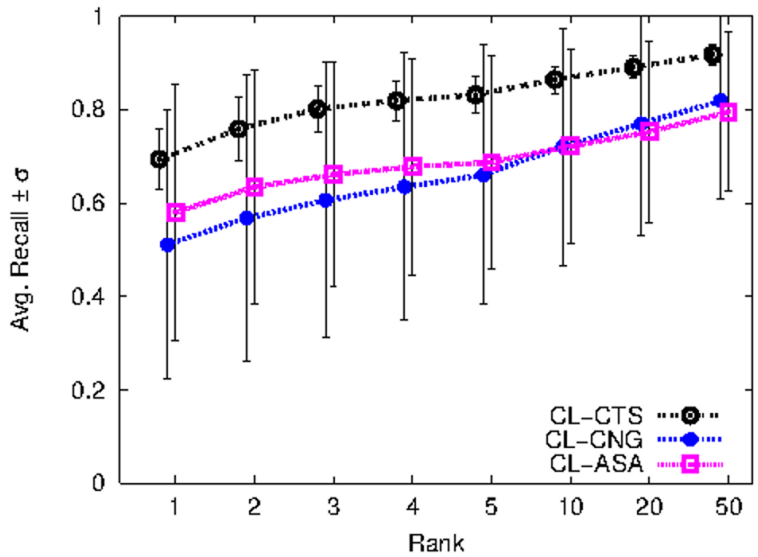
Corpus	Person	Location	Organisation	Total
JRC	1.8%	2.3%	8.7%	12.9%
PAN	1.8%	1.7%	1.9%	5.4%
Wiki	4.7%	3.7%	5.5%	14.0%

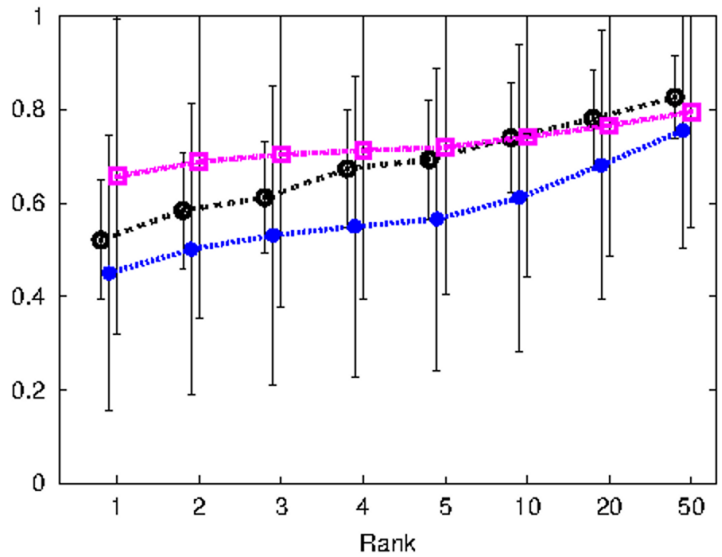
Observations

- CL-ASA performs better on the JRC and very poor on the Wiki
  - better results on nearly parallel data
- CL-CNG performs better on the Wiki and very poor on the PAN
  - better performance on the NE dominated corpora
- CL-CTS exhibits very stable performance across the corpora

Analysis : Average performance and standard deviation

en-es	en-de





Remarks : CL-CTS

- Outperforms
  - char n-gram based model on linguistic corpus (PAN)
  - machine translation based model on comparable corpus (Wiki)
- Achieves a stable performance across the domains using a domain specific thesaurus
- Useful when
  - the nature of data is unknown OR
  - dealing with a heterogeneous data
- Uses reduced concepts and NEs $\rightarrow$ very compact inverted index and low computational cost

Thank You! ☺

Tag cloud of Lab's publications

References I

Alberto Barrón-Cedeño, Paolo Rosso, David Pinto, and Alfons Juan.
2008.
On Cross-lingual Plagiarism Analysis using a Statistical Model.
In *Proceedings of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, PAN'08.

Paul Mcnamee and James Mayfield.
2004.
Character N-Gram Tokenization for European Language Text Retrieval.
Inf. Retr., 7(1-2):73–97, January.

David Pinto, Jorge Civera, Alberto Barrón-Cedeño, Alfons Juan, and Paolo Rosso.
2009.
A Statistical Approach to Crosslingual Natural Language Tasks.
J. Algorithms, 64(1):51–60, January.

Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat.
2006.
Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus.
CoRR, abs/cs/0609059.