Building a Cross-lingual Relatedness Thesaurus using a Graph Similarity Measure

Lukas Michelbacher Florian Laws Beate Dorow Ulrich Heid
Hinrich Schütze

Institute for Natural Language Processing
University of Stuttgart
http://www.ims.uni-stuttgart.de/wiki/extern/WordGraph

LREC 2010 (Valetta, Malta)

May 20, 2010

Overview

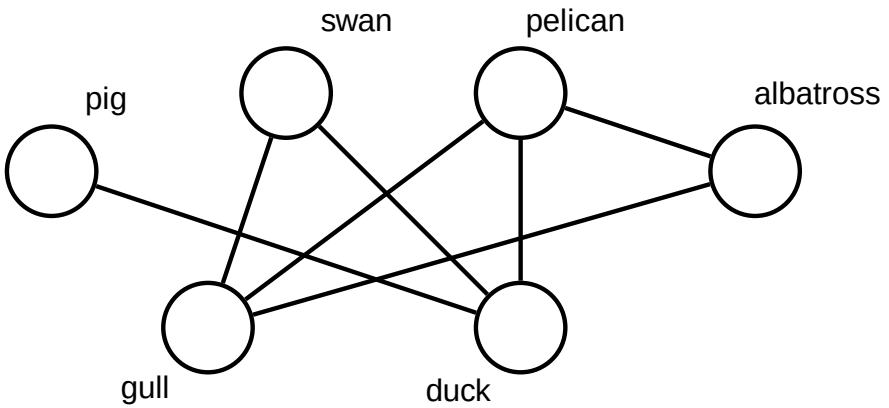
- motivation
- graph similarity measure
- results and evaluation
- related work
- summary
- questions (5 min.)

Motivation

- growing pool of documents
- documents in multiple languages
- need for cross-lingual methods/resources
- **cross-lingual relatedness thesaurus**
- interactive query expansion [Harman, 1988]
- new and open resource
  (`http://www.ims.uni-stuttgart.de/wiki/extern/WordGraph`)

- next: graph similarity measure

Graph Similarity Measure I (Idea)

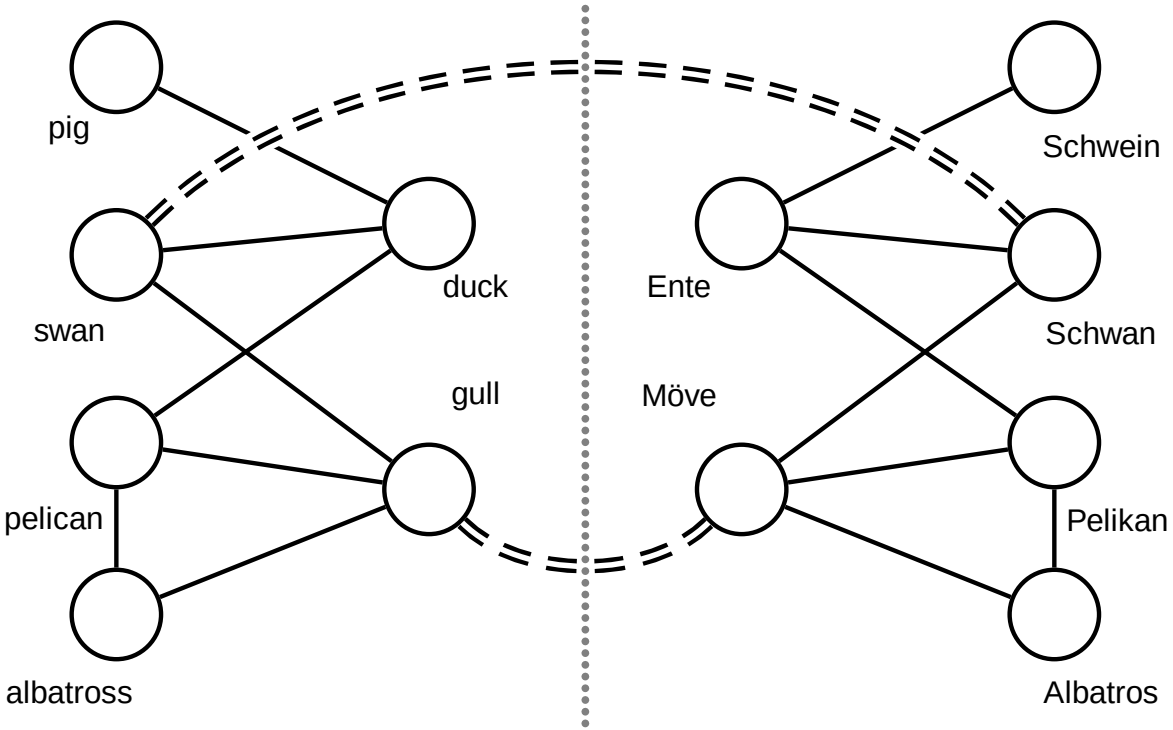


- coordination relation
- original SimRank computation [Jeh and Widom, 2002]

$$S_{ij} = \frac{c}{|N(i)|\ |N(j)|} \sum_{k \in N(i), l \in N(j)} S_{kl}$$

- words are similar if their neighbors are similar, $\forall i : S_{ii} = 1$
- $N(i)$: set of i's neighbors
- c : dampening factor
- similarity spreads throughout the graph with each iteration

Graph Similarity Measure II (Extension)



- SimRank across two graphs $\mathcal{A}$, $\mathcal{B}$ [Dorow et al., 2009]

$$S_{ij} = \frac{c}{|N_{\mathcal{A}}(i)| \; |N_{\mathcal{B}}(j)|} \sum_{k \in N_{\mathcal{A}}(i), l \in N_{\mathcal{B}}(j)} A_{ik} \; B_{jl} \; S_{kl}$$

- self similarity replaced by known translations (“seeds”, dashed lines)
- $S(duck, Ente)$ will benefit from seeds

Graph Similarity Measure III (comparison)

- requires seeds (cf. [Hassan and Mihalcea, 2009])
- does not require aligned corpora (cf. [Sheridan et al., 1997])
- works with different relations
- extendable (see Future Work)

Setup

- two monolingual corpora (German and English Wikipedia, POS-tagged)
- graph representation: words (nodes), relationships (links)
- noun coordinations (lemmas): e.g. *brothers, sisters or other relatives*
- seeds, `dict.cc`
- SimRank [Jeh and Widom, 2002], cross-lingual extension [Dorow et al., 2009]
- based on translations, discover new related words
- next: example, evaluation

Example: thesaurus entry

Ten most related words for test pair *stomach* and *Magen*

rank	related word
1	Magen (*stomach*)
2	Milz (*spleen*)
3	Niere (*kidney*)
4	Leber (*liver*)
5	Kinn (*chin*)
6	Lunge (*lung*)
7	Darm (*bowel*)
8	Bauch (*abdomen*)
9	Gehirn (*brain*)
10	Wange (*cheek*)

rank	related word
1	stomach
2	bladder
3	pancreas
4	spleen
5	colon
6	kidney
7	liver
8	lung
9	duodenum
10	marrow

- next: evaluation

Evaluation I – test set creation

- basis: [Rapp, 1999], 100 word test (automatic lexicon extraction EN$\rightarrow$DE)
- for nouns in test set: manually rated top 10 related words in the thesaurus
- 3 students (2 German native speakers, 1 German-English bilingual)
- categories: **(C)**ohyponyms, hype**(R)**nyms, **(H)**ypohyms, **(E)**xact translations, **(O)**ther
- category **(O)**: semantic relations not covered by the other categories:
  - e.g. *moon – galaxy*, *man – manhood*
- use of English-German dictionary
- agreement: .57 EN $\rightarrow$ DE, .49 for DE $\rightarrow$ EN, Cohen's κ [Artstein and Poesio, 2008]

Evaluation II – results

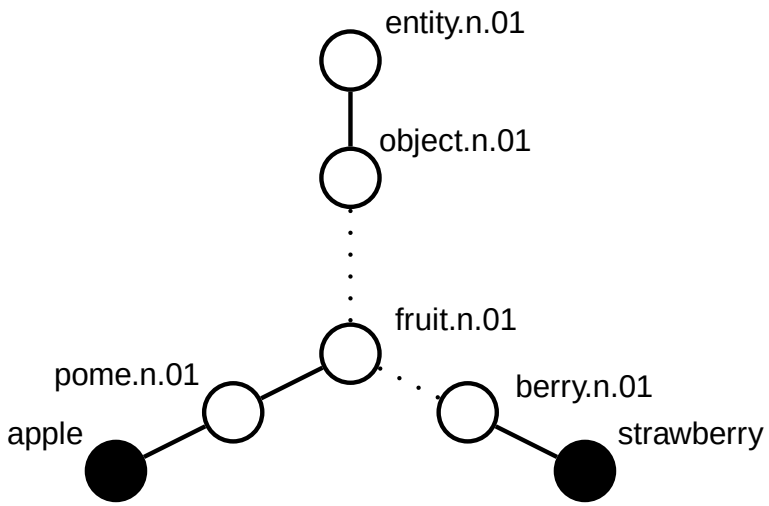
	DE → EN	EN → DE
(C) cohyponyms	28%	22%
(O) other	15%	11%
(E) exact	7%	8%
(H) hyponyms	5%	5%
(R) hypernyms	2%	3%
total related	**57%**	**49%**
unrelated	43%	51%

- cohyponyms dominate

without **(O)**

- performance decreases: 43% (DE → EN) and 39% (EN → DE)
- agreement increases: .62 for EN → DE (.57) and .54 DE → EN (.49)
- next: cohyponym check

Cohyponym check



- spot trivial cohyponyms
- lowest common subsuming hypernym (LCS) in WordNet
- average path length 5
- WordNet coverage
- trivial cohyponyms not a problem
- next: future work

Future Work

- other linguistic relations (“multi-edge”)
- context information
- use thesaurus in IR system
- next: related work

Related Work

- [Hassan and Mihalcea, 2009], cross-lingual semantic relatedness
  - vector-based approach, ESA [Gabrilovich and Markovitch, 2007]
  - Wikipedia inter-language links map concept vectors
- [Sheridan et al., 1997] cross-language similarity thesaurus
  - origin: monolingual query expansion, computed on the index, collection-dependent
  - cross-language: requires aligned corpora
  - not freely available
- [Hsu et al., 2008], cross-lingual query expansion
  - uses online translation services, Wikipedia inter-language links and anchor text
  - two-stage process (translation, expansion)
- [Baroni et al., 2009], corpus-based semantic model
  - aims at inducing concepts, their properties and a conceptual hierarchy
  - monolingual
- next: summary

Summary

- ever growing number of documents in different languages
- method for cross-lingual semantic relatedness
- new resource: relatedness thesaurus
- graph-based word similarity measure
- evaluation (rating experiment)
- 57% semantically related words among (DE $\rightarrow$ EN)

- next: questions


Artstein, R. and Poesio, M. (2008).
Inter-coder agreement for computational linguistics.
Computational Linguistics, 34(4).


Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2009).
Strudel: A corpus-based semantic model based on properties and types.
Cognitive Science.


Dorow, B., Laws, F., Michelbacher, L., Scheible, C., and Utt, J. (2009).
A graph-theoretic algorithm for automatic extension of translation lexicons.
In *EACL 2009 Workshop on Geometrical Models of Natural Language Semantics*.


Gabrilovich, E. and Markovitch, S. (2007).
Computing semantic relatedness using wikipedia-based explicit semantic analysis.
In *IJCAI 2007*.


Harman, D. (1988).

Towards interactive query expansion.
In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 321–331.


Hassan, S. and Mihalcea, R. (2009).
Cross-lingual semantic relatedness using encyclopedic knowledge.
In *EMNLP 2009*. Association for Computational Linguistics.


Hsu, C.-C., Li, Y.-T., Chen, Y.-W., and Wu, S.-H. (2008).
Query expansion via link analysis of wikipedia for clir.
In *7th NTCIR Workshop*, Tokyo, Japan.


Jeh, G. and Widom, J. (2002).
Simrank: A measure of structural-context similarity.
In *KDD '02*, pages 538–543.


Rapp, R. (1999).
Automatic identification of word translations from unrelated English and German corpora.
In *COLING 1999*.


Sheridan, P., Braschler, M., and Schäuble, P. (1997).
Cross-language information retrieval in a multilingual legal domain.
In *ECDL '97: Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, pages 253–268.