Exploring Knowledge Bases for Similarity

Eneko Agirre[‡], **Montse Cuadros**[*] German Rigau[‡], Aitor Soroa[‡]

[‡] IXA NLP Group, University of the Basque Country, Donostia, Basque Country,
e.agirre@ehu.es, german.rigau@ehu.es, a.soroa@ehu.es
[*] TALP center, Universitat Politècnica de Catalunya, Barcelona, Catalonia, cuadros@lsi.upc.edu

LREC Conference, 19 May 2010

Outline

Introduction I

Measuring semantic similarity and relatedness between terms is an important problem in lexical semantics [Budanitsky and Hirst, 2006].

- automobile - car : 3.92

Is used in tasks such as:

- Textual Entailment
- Word Sense Disambiguation
- Information Extraction

Use information in WordNet for finding relation between words / senses

- Paths in WordNet
- Most common subsumer
- Lesk

Introduction II

The techniques used to solve this problem rely on:

- **Pre-existing knowledge resources** (thesauri, semantic networks, taxonomies or encyclopedias) [Alvarez and Lim, 2007, Yang and Powers, 2005, Hughes and Ramage, 2007, Agirre et al., 2009]
- **Distributional properties of words from corpora** [Sahami and Heilman, 2006, Chen et al., 2006, Bollegala et al., 2007, Agirre et al., 2009].
- **Graph-based method** [Hughes and Ramage, 2007]
  - Obtain probability distribution for word in WordNet (probability of concept to be closely related to word)
  - Compute similarity of two probability distributions

Introduction III

[Hughes and Ramage, 2007]

- Random walk algorithm over WordNet,
- Good results on a similarity dataset.

[Agirre et al., 2009]

- Improved *[Hughes and Ramage, 2007]* results
- Provided the best results among WordNet-based algorithms on the Wordsim353 dataset. (comparable to a distributional method over four billion documents)

Outline

Graph-based Similarity

Steps:

1. Represent LKB (e.g. WordNet 1.6) as a graph:
   - Nodes represent concepts $(109, 359)$
   - Edges represent relations
     - Of several types (lexico-semantic, coocurrence etc.)
     - May have some weight attached
     - Can use all relations in WordNet (incl. gloss relations $620, 396$)
     - Undirected links (most of WordNet links have an inverse version)
2. Given word, compute probability distribution over WordNet concepts
3. Given two words, compute similarity of probability distributions

LKB used I

- We have used the knowledge integrated in the Multilingual Central Repository (MCR)[Atserias et al., 2004] to build the graph. More concretly:
  - English WordNet version 1.6
  - WordNet 1.6, WordNet 2.0 relations mapped to 1.6 synsets,
  - eXtended WordNet relations [Mihalcea and Moldovan, 2001]
  - Selectional Preference relations for subjects and objects of verbs [Agirre and Martinez, 2002] (from SemCor)
  - Semantic Coocurrence relations (from SemCor)

LKB used II

We have tried three main versions of the Multilingual Central Repository (MCR)[Atserias et al., 2004] in our experiments to built the graph:

mcr16.all: all relations in the MCR are used, including SemCor related relations.

mcr16.all_wout_sc: all relations except semantic cooccurrence relations.

mcr16.all_wout_semcor: all relations except semantic cooccurrences and selectional preferences.

LKB used III

WordNet 3.0

- wn30: all relations in WordNet 3.0.
- wn30g: all relations in WordNet 3.0, plus the relation between a synset and the disambiguated words in its gloss[1]

KnowNet [Cuadros and Rigau, 2008]

- k5: KnowNet-5, obtained by disambiguating only the first five words from each Topic Signature from the WEB (TSWEB).
- k10: KnowNet-10, obtained by disambiguating only the first ten words from each Topic Signature from the WEB (TSWEB).

[1] http://wordnet.princeton.edu/glosstag

WordNet relations and versions

Source	#relations
MCR1.6 all	1,650,110
Princeton WN1.6	138,091
Princeton WN3.0	235,402
Princeton WN3.0 gloss relations	409,099
Selectional Preferences from SemCor	203,546
eXtended WN	550,922
Co-occurring relations from SemCor	932,008
KnowNet-5	231,163
KnowNet-10	689,610

Table: Number of relations between synsets in each resource.

Example Relations

- **WordNet** [Fellbaum, 1998a]

  tree#n#1 –>hyponym–> teak#n#2

- **Extended WordNet** [Mihalcea and Moldovan, 2001]

  teak#n#2 –>gloss–> wood#n#1

- **spSemCor** [Agirre and Martinez, 2002]

  read#v#1 –>tobj–> book#n#1

- **KnowNet** [Cuadros and Rigau, 2008]

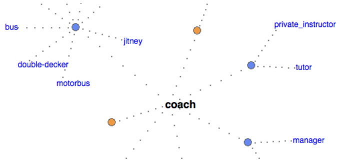
  woodwork#n#2 –>relatedto–> craft#n#1

Outline

UKB

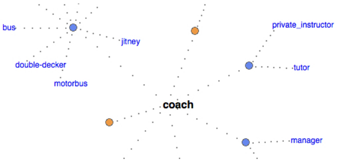
- Set of application for WSD and similarity/relatedness
- Based on graphs
  - Random walks over graphs
  - PageRank and Personalized PageRank
- GPL license
- `http://ixa2.si.ehu.es/ukb/`
- UKB needs three information sources
  - Lexical Knowledge Base (LKB): set of inter-related concepts.
  - Dictionary: link word (lemmas) to LKB concepts.
  - Input context.

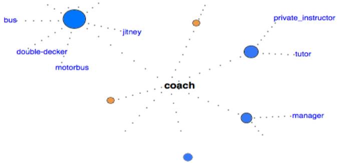
Graph based method



1. Represent LKB (e.g WordNet) as a graph:
   - **Nodes** represent concepts (senses)
   - Undirected **edges** represents semantic relations:
     synonymy, hyperonymy, antonymy, meronymy, entailment, derivation, gloss
2. Apply **PageRank**: Rank nodes (concepts) according to their relative structural importance. Every node has a score.
   - **WSD**: Take best ranked sense of target word
   - **Similarity**: Use the whole vector

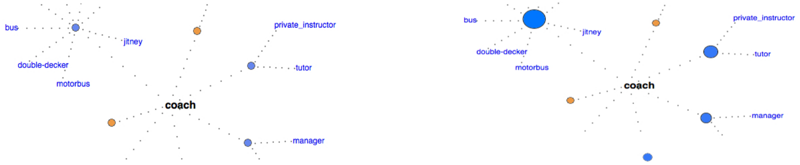
Graph based method





1. Represent LKB (e.g WordNet) as a graph:
   - **Nodes** represent concepts (senses)
   - Undirected **edges** represents semantic relations:
     synonymy, hyperonymy, antonymy, meronymy, entailment, derivation, gloss
2. Apply **PageRank**: Rank nodes (concepts) according to their relative structural importance. Every node has a score.
   - **WSD**: Take best ranked sense of target word
   - **Similarity**: Use the whole vector

Graph based method



1. Represent LKB (e.g WordNet) as a graph:
   - **Nodes** represent concepts (senses)
   - Undirected **edges** represents semantic relations:
     synonymy, hyperonymy, antonymy, meronymy, entailment, derivation, gloss
2. Apply **PageRank**: Rank nodes (concepts) according to their relative structural importance. Every node has a score.
   - **WSD**: Take best ranked sense of target word
   - **Similarity**: Use the whole vector

PageRank

- G: graph with N nodes $n_1, \ldots, n_N$
- d_i: outdegree of node i
- M: $N \times N$ matrix

$$M_{ji} = \begin{cases} \frac{1}{d_i} & \text{an edge from } i \text{ to } j \text{ exists} \\ 0 & \text{otherwise} \end{cases}$$

PageRank equation:

$$\mathbf{Pr} = c M \mathbf{Pr} + (1 - c)\mathbf{v}$$

- voting scheme
- a surfer randomly jumping to any node without following any paths on the graph

c: damping factor: the way in which these two terms are combined at each step

PageRank

- G: graph with N nodes $n_1, \ldots, n_N$
- d_i: outdegree of node i
- M: $N \times N$ matrix

$$M_{ji} = \begin{cases} \frac{1}{d_i} & \text{an edge from } i \text{ to } j \text{ exists} \\ 0 & \text{otherwise} \end{cases}$$

PageRank equation:

$$\mathbf{Pr} = c M \mathbf{Pr} + (1 - c)\mathbf{v}$$

- voting scheme
- a surfer randomly jumping to any node without following any paths on the graph

c: damping factor: the way in which these two terms are combined at each step

PageRank

- G: graph with N nodes $n_1, \ldots, n_N$
- d_i: outdegree of node i
- M: $N \times N$ matrix

$$M_{ji} = \begin{cases} \frac{1}{d_i} & \text{an edge from } i \text{ to } j \text{ exists} \\ 0 & \text{otherwise} \end{cases}$$

PageRank equation:

$$\mathbf{Pr} = cM\mathbf{Pr} + (1-c)\mathbf{v}$$

- voting scheme
- a surfer randomly jumping to any node without following any paths on the graph

c: damping factor: the way in which these two terms are combined at each step

PageRank

- G: graph with N nodes $n_1, \ldots, n_N$
- d_i: outdegree of node i
- M: $N \times N$ matrix

$$M_{ji} = \begin{cases} \frac{1}{d_i} & \text{an edge from } i \text{ to } j \text{ exists} \\ 0 & \text{otherwise} \end{cases}$$

PageRank equation:

$$\mathbf{Pr} = \boxed{cM\mathbf{Pr}} + \boxed{(1-c)\mathbf{v}}$$

- voting scheme
- a surfer randomly jumping to any node without following any paths on the graph

c: damping factor: the way in which these two terms are combined at each step

PageRank

- G: graph with N nodes $n_1, \ldots, n_N$
- d_i: outdegree of node i
- M: $N \times N$ matrix

$$M_{ji} = \begin{cases} \frac{1}{d_i} & \text{an edge from } i \text{ to } j \text{ exists} \\ 0 & \text{otherwise} \end{cases}$$

PageRank equation:

$$\mathbf{Pr} = cM\mathbf{Pr} + (1-c)\mathbf{v}$$

- voting scheme
- a surfer randomly jumping to any node without following any paths on the graph

c: damping factor: the way in which these two terms are combined at each step

Personalized PageRank

$$\mathbf{Pr} = cM\mathbf{Pr} + (1 - c)\mathbf{v}$$

- PageRank: $\mathbf{v}$ is a stocastic normalized vector, with elements $\frac{1}{N}$
  - Equal probabilities to all nodes in case of random jumps
- Personalized PageRank, non-uniform $\mathbf{v}$
  - Assign stronger probabilities to certain kinds of nodes
  - Bias PageRank to prefer these nodes
- For ex. if we concentrate all mass on node i
  - All random jumps return to n_i
  - Rank of i will be high
  - High rank of i will make all the nodes in its vicinity also receive a high rank
  - Importance of node i given by the initial $\mathbf{v}$ spreads along the graph

Personalized PageRank

$$\mathbf{Pr} = cM\mathbf{Pr} + (1-c)\mathbf{v}$$

- PageRank: $\mathbf{v}$ is a stocastic normalized vector, with elements $\frac{1}{N}$
  - Equal probabilities to all nodes in case of random jumps
- Personalized PageRank, non-uniform $\mathbf{v}$
  - Assign stronger probabilities to certain kinds of nodes
  - Bias PageRank to prefer these nodes
- For ex. if we concentrate all mass on node i
  - All random jumps return to n_i
  - Rank of i will be high
  - High rank of i will make all the nodes in its vicinity also receive a high rank
  - Importance of node i given by the initial $\mathbf{v}$ spreads along the graph

Computing Similarity

Given:

$$automobile \text{ –> } \boxed{\text{UKB}} \text{ –> } \vec{automobile}$$
$$car \text{ –> } \boxed{\text{UKB}} \text{ –> } \vec{car}$$

We apply **similartity (**$\vec{automobile}$**,**$\vec{car}$**)** where :

$$\begin{aligned}\mathbf{similarity}(\vec{w},\vec{v}) &= \cos(\theta(\vec{w},\vec{v})) \\ &= \frac{\vec{w}\cdot\vec{v}}{\|\vec{w}\|\|\vec{v}\|} \\ &= \frac{\sum_{i=1}^{n} w_i v_i}{\sqrt{\sum_{i=1}^{n} w_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}\end{aligned}$$

Outline

Definition

Various sets of relations on the WordSim353 dataset [Finkelstein et al., 2002]

tiger, cat
book, paper
computer, keyboard
bread, butter
....

- which contains 353 word pairs, each associated with an average of 13 to 16 human judgements
- Similarity and relatedness are annotated without any distinction.
- Spearman correlation is calculated between gold Standard (WordSim353 dataset) and Similarity probability distribution.

Results

Method	Spearman	Known-words	interval
mcr16.all	0.369690	0.395788	[0.275818, 0.456578]
mcr16.all_wout_sc	0.449606	0.479641	[0.362092, 0.529263]
mcr16.all_wout_semcor	0.525343	0.559497	[0.445263, 0.597086]
mcr16.all_wout_semcor+k5	0.553766	0.589597	[0.476836, 0.622276]
mcr16.all_wout_semcor+k10	0.565809	0.602374	[0.490275, 0.632907]
wn30	0.559087	0.588069	[0.482770, 0.626976]
wn30g	0.658218	0.692505	[0.594597, 0.713647]
wn30g+k5	**0.685184**	**0.720859**	[0.625450, 0.736934]
wn30g+k10	0.638901	0.672213	[0.572612, 0.696891]

Comparision with previous work

Method	**Source**	**Spearman**
[Agirre et al., 2009]	Combination	**0.78**
[Gabrilovich and Markovitch, 2007]	Wikipedia	0.75
This work	WordNet	0.69
[Agirre et al., 2009]	WordNet	0.66
[Agirre et al., 2009]	Web Corpus	0.65
[Gabrilovich and Markovitch, 2007]	ODP	0.65
[Finkelstein et al., 2002]	Combination	0.56
[Finkelstein et al., 2002]	LSA	0.56
[Hughes and Ramage, 2007]	WordNet	0.55

Outline

Conclusions

The main conclusions from the results are the following:

- The best combinations for MCR1.6 are obtained ignoring selectional preferences and semantic occurrences.
- The disambiguated glosses improve the results by a large margin on wn30.
- KnowNet improves results in both datasets. The largest gains are for MCR1.6 with KnowNet-10 (k10), but the best overall results are for Wordnet3.0 with disambiguated glosses and KnowNet-5 (k5)
- Results show that using the adequate relations the performance improves over previously published WordNet-based results on the WordSim353 dataset.
- Similarity software and some graphs used in this paper are publicly available at `http://ixa2.si.ehu.es/ukb`

Future Work

- Similar study on WSD using a related algorithm[Agirre and Soroa, 2009],
- Compare which is the best setting on these closely interrelated tasks.

Exploring Knowledge Bases for Similarity

Eneko Agirre[‡], **Montse Cuadros**[*] German Rigau[‡], Aitor Soroa[‡]

[‡] IXA NLP Group, University of the Basque Country, Donostia, Basque Country,
e.agirre@ehu.es, german.rigau@ehu.es, a.soroa@ehu.es
[*] TALP center, Universitat Politècnica de Catalunya, Barcelona, Catalonia, cuadros@lsi.upc.edu

LREC Conference, 19 May 2010

Agirre, E. and Lopez de Lacalle, O. (2004).
Publicly available topic signatures for all wordnet nominal senses.
In *Proceedings of LREC*.

Agirre, E. and Martinez, D. (2002).
Integrating selectional preferences in wordnet.
In *Proceedings of the First International WordNet Conference*, Mysore, India.

Agirre, E. and Soroa, A. (2008).
Using the multilingual central repository for graph-based word sense disambiguation.
In *Proceedings of LREC*.

Agirre, E. and Soroa, A. (2009).
Personalizing pagerank for word sense disambiguation.
In *Proc. of EACL 2009*, Athens, Greece.

Agirre, E., Soroa, A., Alfonseca, E., Hall, K., Kravalova, J., and Pasca, M. (2009).
A study on similarity and relatedness using distributional and WordNet-based approaches.

In *Proceedings of annual meeting of the North American Chapter of the Association of Computational Linguistics (NAAC)*, Boulder, USA.

Alvarez, M. and Lim, S. (2007).
A graph modeling of semantic similarity between words.
Proceedings of the Conference on Semantic Computing, pages 355–362.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., and Vossen, P. (2004).
The meaning multilingual central repository.
In *Proc. of Global WordNet Conference*, Brno, Czech Republic.

Bollegala, D., Y., M., and Ishizuka, M. (2007).
Measuring semantic similarity between words using web search engines.
In *Proceedings of WWW'2007*.

Budanitsky, A. and Hirst, G. (2006).
Evaluating WordNet-based Measures of Lexical Semantic Relatedness.
Computational Linguistics, 32(1):13–47.

Chen, H., Lin, M., and Wei, Y. (2006).
Novel association measures using web search with double checking.
In *Proceedings of COCLING/ACL 2006*.

Cuadros, M. and Rigau, G. (2008).
KnowNet: Building a Large Net of Knowledge from the Web.
In *Proceedings of COLING*.

Cuadros, M., Rigau, G., and Castillo, M. (2007).
Evaluating large-scale knowledge resources across languages.
In *Proceedings of RANLP*.

Daudé, J., Padró, L., and Rigau, G. (2003).
Making Wordnet Mappings Robust.
In *Proceedings of the 19th Congreso de la Sociedad Espala para el Procesamiento del Lenguage Natural, SEPLN'03*, Universidad Universidad de Alcala de Henares. Madrid, Spain.

Fellbaum, C. (1998a).
WordNet. An Electronic Lexical Database.
Language, Speech, and Communication. The MIT Press.

Fellbaum, C., editor (1998b).
WordNet: An Electronic Lexical Database and Some of its Applications.
MIT Press, Cambridge, Mass.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002).
Placing Search in Context: The Concept Revisited.
ACM Transactions on Information Systems, 20(1):116–131.

Gabrilovich, E. and Markovitch, S. (2007).
Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis.
Proc of IJCAI, pages 6–12.

Haveliwala, T. H. (2002).
Topic-sensitive pagerank.
In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526.

Hughes, T. and Ramage, D. (2007).
Lexical semantic relatedness with random graph walks.
In *Proceedings of EMNLP-CoNLL-2007*, pages 581–589.

Laparra, E. and Rigau, G. (2009).
Integrating wordnet and framenet using a knowledge-based word sense disambiguation algorithm.

In *Proceedings of Recent Advances in Natural Language Processing (RANLP09)*, Borovets, Bulgaria.

Mihalcea, R. and Moldovan, D. (2001).
extended wordnet: Progress report.
In *NAACL Workshop* WordNet and Other Lexical Resources: Applications, Extensions and Customizations *(NAACL'2001).*, pages 95–100, Pittsburg, PA, USA.

Miller, G. and Charles, W. (1991).
Contextual correlates of semantic similarity.
Language and Cognitive Processes, 6(1):1–28.

Miller, G., Leacock, C., Tengi, R., and Bunker, R. (1993).
A Semantic Concordance.
In *Proceedings of the ARPA Workshop on Human Language Technology*.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999).
The pagerank citation ranking: Bringing order to the web.
Technical Report 1999-66, Stanford InfoLab.
Previous number = SIDL-WP-1999-0120.

Resnik, P. (1995).

Using Information Content to Evaluate Semantic Similarity in a Taxonomy.
Proc. of IJCAI, 14:448–453.

Sahami, M. and Heilman, T. (2006).
A web-based kernel function for measuring the similarity of short text snippets.
Proc. of WWW, pages 377–386.

Yang, D. and Powers, D. (2005).
Measuring semantic similarity in the taxonomy of WordNet.
Proceedings of the Australasian conference on Computer Science.