Highlighting relevant concepts from Topic Signatures

Montse Cuadros*, Lluís Padró⊙, German Rigau◇

* Vicomtech-IK4, Donostia-San Sebastián mcuadros@vicomtech.org
⊙ TALP center, Universitat Politècnica de Catalunya, Barcelona, padro@lsi.upc.edu
◇ IXA NLP Group, University of the Basque Country, Donostia-San Sebastián, german.rigau@ehu.es

Abstract

This paper presents deepKnowNet, a new fully automatic method for building highly dense and accurate knowledge bases from existing semantic resources. Basically, the method applies a knowledge-based Word Sense Disambiguation algorithm to assign the most appropriate WordNet sense to large sets of topically related words acquired from the web, named TSWEB. This Word Sense Disambiguation algorithm is the personalized PageRank algorithm implemented in UKB. This new method improves by automatic means the current content of WordNet by creating large volumes of new and accurate semantic relations between synsets. KnowNet was our first attempt towards the acquisition of large volumes of semantic relations. However, KnowNet had some limitations that have been overcomed with deepKnowNet. deepKnowNet disambiguates the first hundred words of all Topic Signatures from the web (TSWEB). In this case, the method highlights the most relevant word senses of each Topic Signature and filter out the ones that are not so related to the topic. In fact, the knowledge it contains outperforms any other resource when is empirically evaluated in a common framework based on a similarity task annotated with human judgements.

Keywords: Knowledge Acquisition, Word Sense Disambiguation, Lexical Semantics

1. Introduction

Using large-scale knowledge bases, such as WordNet (Fellbaum, 1998), has become a usual, often necessary, practice for most current Natural Language Processing (NLP) systems. Even now, building large and rich enough knowledge bases for broad–coverage semantic processing takes a great deal of expensive manual effort involving large research groups during long periods of development. In fact, hundreds of person-years have been invested in the development of wordnets for various languages (Vossen, 1998). For example, in more than ten years of manual construction (from 1995 to 2006, that is from version 1.5 to 3.0), WordNet grew from 103,445 to 235,402 semantic relations[1]. But this data does not seem to be rich enough to support advanced concept-based NLP applications directly. It seems that applications will not scale up to working in open domains without more detailed and rich general-purpose (and also domain-specific) semantic knowledge built by automatic means. Obviously, this fact has severely hampered the state-of-the-art of advanced NLP applications.

However, the Princeton WordNet (WN) is by far the most widely-used knowledge base (Fellbaum, 1998). In fact, WordNet is being used world-wide for anchoring different types of semantic knowledge including wordnets for languages other than English (Atserias et al., 2004), domain knowledge (Magnini and Cavaglià, 2000) or ontologies like SUMO (Niles and Pease, 2001) or the EuroWordNet Top Concept Ontology (Àlvez et al., 2008). It contains manually coded information about nouns, verbs, adjectives and adverbs in English and is organized around the notion of a *synset*. A synset is a set of words with the same part-of-speech that can be interchanged in a certain context. For example, <*party*, *political_party*> form a synset because they can be used to refer to the same concept. A synset is often further described by a gloss, in this case: "an organization to gain political power" and by explicit semantic relations to other synsets.

Moreover, during the last years the research community has devised a large set of innovative methods and tools for large-scale automatic acquisition of lexical knowledge from structured and unstructured corpora. Among others we can mention the acquisition of Selectional Preferences from raw corpora (Resnik, 1993) or SemCor (Agirre and Martínez, 2002), synonyms (Lin and Pantel, 2002), sense examples (Leacock et al., 1998), knowledge about individuals from Wikipedia (Suchanek et al., 2007), topic signatures from corpora (Lin and Hovy, 2000), the web (Agirre and Lopez de Lacalle, 2004), new concepts and relations between concepts (Moldovan and Gîrju, 2000), paraphrases (Lin and Pantel, 2001), sense frequencies (McCarthy et al., 2004), co-occurrence feature vectors for WordNet synsets (Pantel, 2005), hyponymy relations (Snow et al., 2006).

Obviously, all these semantic resources have been acquired using a very different set of processes, tools and corpora. As expected, each semantic resource has different volume and accuracy figures when evaluated in a common and controlled framework (Cuadros and Rigau, 2006).

However, not all these large-scale resources encode semantic relations between synsets. In some cases, only relations between synsets and words have been acquired. This is the case of the Topic Signatures acquired from the web (Agirre and Lopez de Lacalle, 2004). This is one of the largest semantic resources ever built with around one hundred million relations between synsets and semantically related words [2].

A knowledge net or KnowNet (KN)[3] (Cuadros and Rigau, 2008), is an extensible, large and accurate knowledge base, which has been derived by semantically disambiguating the

[1]Symmetric relations are counted only once.

[2]http://ixa.si.ehu.es/Ixa/resources/sensecorpus

[3]http://adimen.si.ehu.es/web/KnowNet

Topic Signatures acquired from the web(TSWEB) (Agirre et al., 2001, Agirre and Lopez de Lacalle, 2004). Basically, the method uses a robust and accurate knowledge-based Word Sense Disambiguation algorithm to assign the most appropriate senses to the topic words associated to a particular synset. The resulting knowledge-base which connects large sets of topically-related concepts is a major step towards the autonomous acquisition of knowledge from raw text.

KnowNets performs similarly to other knowledge bases developed by manual or automatic means in multiple evaluation scenarios, they directly perform better than many other existing knowledge bases. However, it seems that the intrinsic density of KnowNets does not help in some other scenarios. Moreover, KnowNet only use at most the first twenty words from every topic signature when on average they contain more than a thousand of words. Obviously, many relevant concepts appear in a topic signature beyond the first twenty words.

For instance, Table 1 presents the first hundred words of the topic signature acquired from the web (TSWEB) corresponding to horse^1_n. In this case, words are ordered by their relevance weight. This word sense is described as *"solid-hoofed herbivorous quadruped domesticated since prehistoric times"* in WordNet 1.6.

Most of the first twenty words from this topic signature correspond to different types of $horse_n$ or closely related equids except $polo_n$ which corresponds to a game where horses play a role, $liver_n$ which corresponds to an organ, $equid_a$ which does not exist in WordNet (obviously an error of the POS tagger) and $mussel_n$ a marine bivalve (obviously an error). Those concepts closely related to equids appear because of the intrinsic way topic signatures are built. TSWEB use closely related monosemous relatives of the target topic to build a query for a web sarch engine. Obviously, these terms appear more frequently in the subcorpus obtained from that query and thus, these words appear higher in the ranking.

However, many words relevant to $horse_n$ appear beyond the first twenty words. For instance, $saddle_n$, $race_n$, $equid_n$, $riding_n$, $stable_n$, etc.

Most of the first twenty words from this topic signature correspond to different types of $horse_n$ or closely related equids except $polo_n$ which corresponds to a game where horses play a role, $liver_n$ which corresponds to an organ, $equid_a$ which does not exist in WordNet (obviously an error of the POS tagger) and $mussel_n$ a marine bivalve (obviously another error). Those concepts closely related to equids appear because of the intrinsic way topic signatures are built. TSWEB use closely related monosemous relatives of the target topic to build a query for a web search engine. Obviously, these terms appear more frequently in the subcorpus obtained from that query and thus, these words appear higher in the ranking.

Obviously, TSWEB contain misleading pitfalls and deficiencies. Basically, we can mention:

- Basic linguistic preprocessing errors (i.e. tokenization, lemmatization, POS tagging).
- Misleading words appearing due to the acquisition method including a proper characterization of the topic query.
- Misleading weights for the relevant words due to the examples retrieved and the formulas applied.

Trying to avoid these shortcomings, this paper explores a new method for building KnowNets. The method locates the most relevant concepts of a Topic Signature by applying a graph-based similarity algorithm. Those concepts are selected for building a new resource, we call deepKnowNet. This paper is organized as follows. Firstly, Section 2. presents the method used to build deepKnowNets. Secondly, Section 3. presents the preliminary evaluation of the new knowledge bases and finally, Section 4. presents some conclusions and a preliminary future research line.

2. Building deepKnowNet

We apply a new approach for building KnowNets of a better quality called deepKnowNets. Basically, the new method explores the first hundred words of every topic signature, and selects the most relevant concepts according to a similarity measure provided by UKB[4] (Agirre and Soroa, 2009) on an *initial* knowledge base consisting on WordNet (Fellbaum, 1998) and eXtendedWN (Mihalcea and Moldovan, 2001).

Figure 1 presents the basic schema we follow to automatically acquire deepKnowNets.

- For each **sense** (topic) having an associated topic signature (*ts*).
  - **step 1:** Obtain the *personalized PageRank vector* (*ppv*) for the given *sense* applying an algorithm included in UKB by using a knowledge base consisting on WordNet (Fellbaum, 1998) and eXtendedWN (Mihalcea and Moldovan, 2001).
  - **step 2:** Apply a *sorting-filtering* process:
    - ∗ Sort *ppv* with respect the similarity weight to the given *sense* (topic).
    - ∗ Filter out from *ppv* the senses of the words not appearing also in the first hundred positions of *ts*.
    - ∗ Select a subset of concepts from *ppv* to build the final deepKnowNets.
  - **step 3:** Create a deepKnowNet with the semantic relations obtained from the previous step. We directly connect the *sense* (topic) to the selected senses from *ppv*.

    To create a deepKnowNet we used two different methods to combine the word-senses and obtain the knowledge bases. They are *Direct relations* and *all-with-all*.

    Direct relations connects the word sense from the topic with every disambiguated term from the topic signature. For instance, consider the set of senses from the disambiguated topic signature for party^1_n:

[4] `http://ixa2.si.ehu.es/ukb/`

polo$_n$ equus$_n$ zebra$_n$ eohippus$_n$ quagga$_n$ horse$_n$ pony$_n$ hinny$_n$ caballus$_n$ stablemate$_n$ racehorse$_n$ donkey$_n$ liver$_n$ mare$_n$ equid$_a$ mussel$_n$ pinto$_n$ bangtail$_n$ workhorse$_n$ palomino$_n$ stallion$_n$ **saddle**$_n$ dawn$_n$ **race**$_n$ mesohippus$_n$ **equid**$_n$ **riding**$_n$ companion$_n$ harness$_n$ specie$_n$ extinct$_a$ offspring$_n$ chestnut$_n$ hyracotherium$_n$ **ride**$_v$ ass$_n$ ancestor$_n$ female$_a$ male$_a$ filly$_n$ foal$_n$ stable$_a$ trainer$_n$ fossil$_n$ mule$_n$ race$_v$ female$_n$ dreissena$_n$ asinus$_n$ burro$_n$ thoroughbred$_a$ Thoroughbred$_v$ hybrid$_n$ breeding$_n$ racing$_n$ modern$_a$ champion$_n$ ago$_r$ own$_v$ age$_n$ broodmare$_n$ finch$_n$ mammal$_n$ breed$_n$ dog$_n$ printer$_n$ breed$_v$ colt$_n$ wild$_n$ hybrid$_a$ owner$_n$ equine$_n$ Gee-Gee$_a$ Przewalski$_n$ bugensis$_n$ derby$_n$ foal$_a$ midget$_n$ oligocene$_n$ sterile$_a$ ownership$_n$ arabian$_n$ genus$_n$ domestic$_a$ **stable**$_n$ wild$_a$ Breyer$_n$ Standardbred$_v$ eocene$_n$ mustang$_n$ subspecies$_n$ trail$_n$ animal$_n$ bean$_n$ sire$_n$ stud$_n$ gelding$_n$ polymorpha$_n$ sheep$_n$ evolution$_n$

Table 1: First hundred words of horse^1_n from TSWEB ordered by its relevance weight

tamany^1_n
alinement^1_n
greenback^1_n
constitutional^1_n
federalist^1_n
whig^3_n
nazi^1_a
republican^1_n

We select all pairs that combine party^1_n with the rest of senses of the topic signature:

party^1_n related-to tamany^1_n
party^1_n related-to alinement^1_n
party^1_n related-to greenback^1_n
party^1_n related-to constitutional^1_n
party^1_n related-to federalist^1_n
party^1_n related-to whig^3_n
party^1_n related-to nazi^1_a
party^1_n related-to republican^1_n

The second method, *all-with-all*, produces much dense knowledge bases. This method creates a new relation for each possible pair of senses in the disambiguated topic signature. Using this method in the previous example, we would obtain the following relations:

party^1_n related-to tamany^1_n
party^1_n related-to alinement^1_n
party^1_n related-to greenback^1_n
party^1_n related-to constitutional^1_n
party^1_n related-to federalist^1_n
party^1_n related-to whig^3_n
party^1_n related-to nazi^1_a
party^1_n related-to republican^1_n

tamany^1_n related-to alinement^1_n
tamany^1_n related-to greenback^1_n
tamany^1_n related-to constitutional^1_n
tamany^1_n related-to federalist^1_n
tamany^1_n related-to whig^3_n
tamany^1_n related-to nazi^1_a
tamany^1_n related-to republican^1_n

alinement^1_n related-to greenback^1_n
alinement^1_n related-to constitutional^1_n
alinement^1_n related-to federalist^1_n
alinement^1_n related-to federalist^1_n
alinement^1_n related-to whig^3_n
alinement^1_n related-to nazi^1_a
alinement^1_n related-to republican^1_n

etc.

Note that this new method does not consider the application of an explicit Word Sense Disambiguation algorithm. Instead, the sorting-filtering process removes undesired interpretations.

Table 2 shows the impact of reordering when using *ppv* on the first hundred words from the same topic signature presented in Table 1 for the sense horse^1_n. Obviously, the words in this vector have been reordered with respect the original topic signature.

In bold we present the words appearing in the first twenty positions from Table 1. Note that there are only 18 words in bold[5]. Now, the first twenty positions seem to be related to the topic but only a small subset belong to the set of monosemous relatives. Additionally, some relevant words such as *saddle*$_n$, *race*$_n$, *equid*$_n$, *riding*$_n$, *stable*$_n$, still appear among the most relevant.

We developed several new deepKnowNets[6]. The differences among them correspond to the final set of concepts selected (step 2). In step 2, we select the final set of concepts from a filtered and sorted *ppv* as those concepts covering a percentage of the total weight (i.e. 80%, 85%, 90%, 95% and 99% of the total weight). For instance, Table 3 shows the selected word senses when using 80%, 85%, 90% and 95% of the total weights from *ppv* for sense *party*^1_n. For instance, deepKnowNet-80d refer to a deepKnowNet built using those concepts from a filtered and sorted *ppv* covering 80% of the total weights and using a *direct* combination method.

Table 4 presents the total amount of relations contained in both KnowNets and deepKnowNets. Although they are quite similar in sizes (ranging from hundreds of thousands to millions), deepKnowNets have been developed using the *direct* combination method while KnowNets used the *all with all* combination.

[5] *caballus*$_n$ and *equid*$_a$ do not appear in WordNet
[6] http://adimen.si.ehu.es/web/deepKnowNet

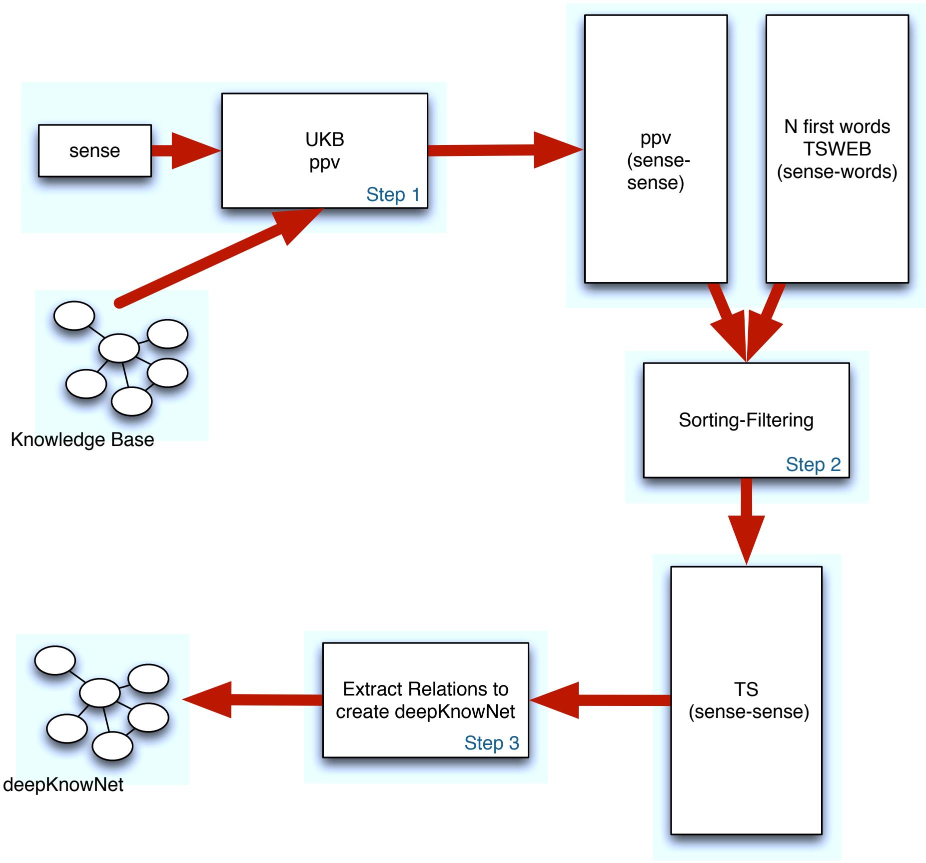

Figure 1: DeepKnowNet schema

age$_n$ racing$_n$ riding$_n$ **polo**$_n$ broodmare$_n$ own$_v$ **zebra**$_n$ gelding$_n$ owner$_n$ arabian$_n$ filly$_n$ specie$_n$ breeding$_n$ colt$_n$ **hinny**$_n$ male$_a$ thoroughbred$_a$ sheep$_n$ mammal$_n$ female$_a$ dog$_n$ race$_v$ trail$_n$ **donkey**$_n$ mule$_n$ extinct$_a$ breed$_n$ harness$_n$ genus$_n$ race$_n$ sire$_n$ **stablemate**$_n$ chestnut$_n$ mesohippus$_n$ **palomino**$_n$ **eohippus**$_n$ domestic$_a$ stable$_n$ mustang$_n$ **pinto**$_n$ hyracotherium$_n$ foal$_n$ **workhorse**$_n$ **equus**$_n$ stud$_n$ **mare**$_n$ stallion$_n$ equid$_n$ equine$_n$ ass$_n$ **pony**$_n$ **racehorse**$_n$ **bangtail**$_n$ saddle$_n$ ride$_v$ breed$_v$ animal$_n$ **horse**$_n$ companion$_n$ trainer$_n$ fossil$_n$ wild$_n$ bean$_n$ finch$_n$ champion$_n$ ago$_r$ hybrid$_n$ evolution$_n$ ownership$_n$ derby$_n$ **liver**$_n$ female$_n$ ancestor$_n$ offspring$_n$ oligocene$_n$ printer$_n$ **mussel**$_n$ subspecies$_n$ dawn$_n$ sterile$_a$ wild$_a$ burro$_n$ dreissena$_n$ modern$_a$ **quagga**$_n$ eocene$_n$ stable$_a$ midget$_n$ hybrid$_a$

Table 2: First set of hundred words of horse^1_n from TSWEB ordered using *ppv*, in this case, 11 are gone because did not appear in the KB (WN+XWN).

3. Evaluation

In order to evaluate the new deepKnowNets, we carried out two experiments. The first one studies the different reordering of words with respect the original order of the topic signatures. The second one compares the accuracy of the previous KnowNets with respect the new deepKnowNets in the same similarity task.

3.1. Re-ranking

The first experiment compares the spearman correlation between the original topic signatures and the new ones. We can illustrate graphically the differences by using some plots.

All figures show in the x-axis the original ordering of TSWEB and in the y-axis the new reordering.

Figure 2 shows at the left hand side the relation between the order of the first twenty words of the topic signature of horse^1_n from TSWEB and its final positions after the reordering process. At the right side, instead of first twenty words from horse^1_n, Figure 2 shows the same data but including the first hundred words from TSWEB. We also plot in both figures line $x = y$ to represent the original order.

Figure 3 shows at the left hand side the relation between the order of the first hundred words of the topic signature of a randomly chosen topic signature[7] and its final positions after the reordering process. At the right side, instead of a single topic signature, Figure 3 shows the same data but

[7] corresponding to epilogue^1_n

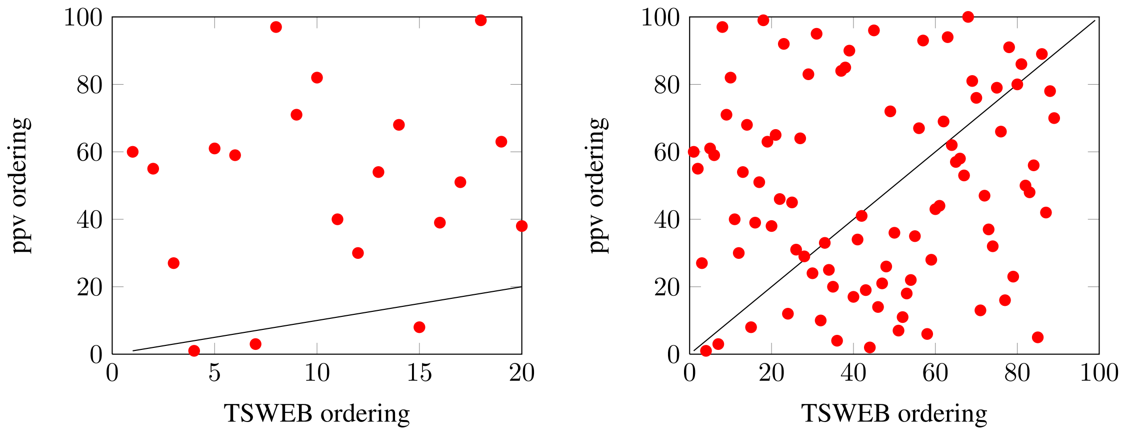

Figure 2: Relative ordering positions of the first twenty (left figure) and hundred (right figure) words of horse^1_n's.

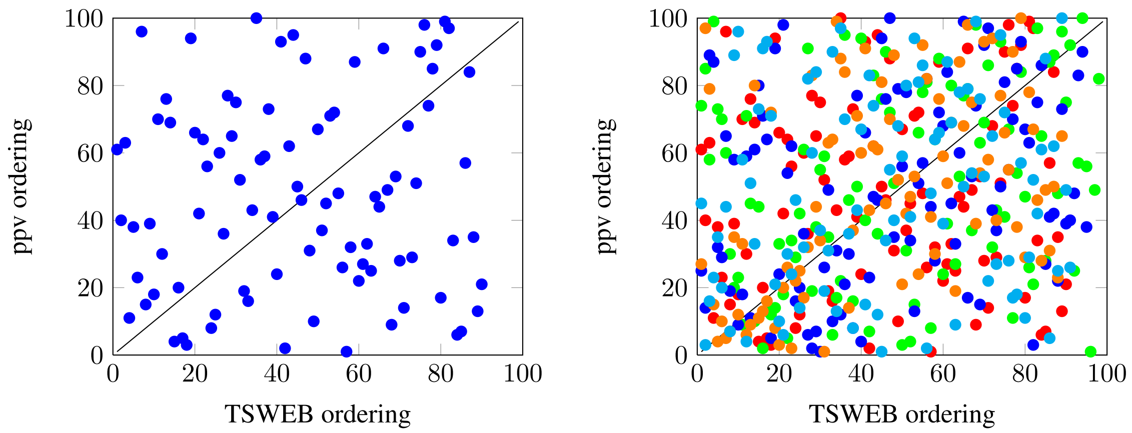

Figure 3: Relative ordering positions of the first hundred words of one (left figure) and five (right figure) random topic signatures.

considering five random topic signatures[8]. We also plot in both figures line $x = y$ to represent the original order.
Additionally, we calculated the spearman correlation between both orderings of the the first hundred words of the topic signature of *horse*^1_n. When this coefficient is close to one, the results are very similar to those of the gold-standard, and when close to zero, the results are very different. For *horse*^1_n, the spearman correlation is 0.096536. Obviously, as the correlation is near to zero, they seem to be different.
Moreover, the mean spearman correlation between both orderings of the first hundred words of the all the topic signatures from TSWEB is only 0.12353. This result indicates that, in general, both orderings are very different.

3.2. Similarity task

We tested the different knowledge bases on the WordSim353 dataset (Finkelstein et al., 2002) [9], which contains 353 word pairs, each associated with an average of 13 to 16 human judgements. In this dataset, both similarity and relatedness are annotated without any distinction. Several studies indicate that the human scores consistently have very high correlations with each other (Miller and Charles, 1991, Resnik, 1995), thus validating the use of these kind of datasets for evaluating semantic similarity.
We use different knowledge bases (i.e. deepKnowNets) to measure the similarity of a word pair. The whole process performs as follows:

1. Calculating *personalized PageRank* vectors for each word using UKB (Agirre and Soroa, 2009) with a graph created from a particular knowledge base (i.e. a Knownet).

2. Obtaining the similarity measure for every word-pair, applying the cosine formula (Equation 1) to both word-pair vectors. A similarity vector is generated by processing all word-pairs of the dataset.

3. Computing the spearman correlation (Equation 2) between the gold-standard and the similarity vector obtained in the previous step.

[8] epilogue^1_n, drift^2_n, expulsion^3_n, ark^2_n, progression^3_n

[9] `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html`

Percentage	Word-sense	Weight
	organization$_n^1$	0.0023064
	politician$_n^2$	0.0022234
	constitute$_v^3$	0.0017662
	political_program$_n^1$	0.0015342
	Labor_Party$_n^1$	0.0015178
	Nazi_n^1	0.0014612
	Federalist$_n^1$	0.0013917
	Republican_Party$_n^1$	0.0013784
	States'_Rights_Democratic_Party$_n^1$	0.0013776
	launch$_v^1$	0.0008944
	elector$_n^1$	0.0005264
	policy$_n^1$	0.0005141
	Communist$_n^1$	0.0004839
	Democrat$_n^1$	0.0004033
	bondage$_n^1$	0.0003372
	Whig$_n^3$	0.0003215
80%	election$_n^1$	0.0002737
	Adolf_Hitler$_n^1$	0.0002584
	opposition$_n^5$	0.0002190
	Tammany_Society$_n^1$	0.0002052
	right-wing$_a^1$	0.0001893
	Republican$_n^1$	0.0001889
85%	Nazi$_a^1$	0.0001842
	liberal$_n^1$	0.0001808
	socialist$_n^1$	0.0001806
	conservative$_a^1$	0.0001784
	American$_a^1$	0.0001511
	democratic$_a^2$	0.0001489
	elect$_v^1$	0.0001483
90%	position$_n^6$	0.0001472
	reform$_n^1$	0.0001396
	structure$_n^1$	0.0001362
	presidential$_a^1$	0.0001345
	political_science$_n^1$	0.0001337
	prohibition$_n^2$	0.0001183
	bull$_n^1$	0.0001102
	federal$_a^2$	0.0001088
	tendency$_n^1$	0.0001069
	Theodore_Roosevelt$_n^1$	0.0001068
	campaign$_n^2$	0.0001026
95%	European_elk$_n^1$	0.0001009

Table 3: Example of the selected word senses when using 80%, 85%, 90% and 95%, of the total weights for sense *party*$_n^1$

Source	#relations
deepKnowNet-80d	203,563
deepKnowNet-85d	263,534
deepKnowNet-90d	356,726
deepKnowNet-95d	549,330
deepKnowNet-99d	1,068,101
KnowNet-5	231,163
KnowNet-10	689,610
KnowNet-15	1,378,286
KnowNet-20	2,358,927

Table 4: Number of synset relations of different KnowNet and deepKnowNet versions

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

$$\rho = \frac{\sum_i (x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_i (x_i-\overline{x})^2 \sum_i (y_i-\overline{y})^2}} \qquad (2)$$

Once performed these steps, we have a correlation coefficient for all word pairs. When this coefficient is close to one, the results are very similar to those of the gold-standard, and when close to zero, the results are very different.

KB	Spearman	Known-words
deepKnowNet-95d	0.597308	0.632745
deepKnowNet-90d	0.595818	0.631181
deepKnowNet-85d	0.590487	0.622690
deepKnowNet-99d	0.581687	0.607556
WN+XWN	0.594069	0.603632
KnowNet-20	0.555765	0.590256
deepKnowNet-80d	0.557507	0.587826
WN	0.568724	0.577710
KnowNet-15	0.537680	0.571447
KnowNet-10	0.485961	0.517732
XWN	0.509375	0.517536
KnowNet-5	0.428597	0.472478

Table 5: Sperman correlation of KnowNets and deepKnowNets in wordSim353 dataset

Table 5 presents the results of the original KnowNets and the new deepKnowNets on the same similarity task. We also include in the comparison other knowledge bases such as WordNet (WN) (Fellbaum, 1998), eXtended WordNet (XWN) (Mihalcea and Moldovan, 2001) and its combination WN+XWN. DeepKnowNets clearly outperform previous versions of KnowNets, WordNet and eXtended WordNet. Additionally, when using more than 80% of the total weights during the *sorting-filtering* step, deepKnowNets also obtain better performances than the *initial* knowledge base necessary for their construction (WN+XWN). As expected, the best results are not obtained when using most of the relations (i.e. deepKnowNet-99d).

Finally, trying to stablish a fair comparison between both KnowNet and deepKnowNet approaches, we developed two additional deepKnowNet versions. The first version, deepKnowNet-20ppvRank has been constructed using the first twenty word senses from the final *ppv* (step 2) and the *all-with-all* combination method (step 3). The second version, deepKnowNet-20tswebRank has been constructed using also the first twenty word senses from the final *ppv* but using the original ranking of TSWEB (step 2) and also the *all-with-all* combination method (step 3). These two new knowledge bases try to be as similiar as possible to KnowNet-20.

KB	Spearman	Known-words
deepKnowNet-20ppvRank	0.583053	0.606265
KnowNet-20	0.555765	0.590256
deepKnowNet-20tswebRank	0.534758	0.559181

Table 7: wordSim353 results

Source	#relations
deepKnowNet-20ppvRank	3,496,860
deepKnowNet-20tswebRank	3,693,978
KnowNet-20	2,358,927

Table 6: Number of synset relations of KnowNet and new deepKnowNet versions

Table 6 presents the total number of semantic relations encoded in KnowNet-20 and the new deepKnowNets.
Table 7 presents the performances of these new deepKnowNets when compared to KnowNet-20. Interestingly, deepKnowNet-20ppvRank obtains the best spearman coefficient similarity measure. This comparison clearly indicates that reordering using *personalized PageRank* positively impacts the resulting knowledge bases. In fact, using the same approach deepKnowKnet-20ppvRank clearly surpass deepKnowNet-20tswebRank. Furthermore, it also seems that this new method for building automatically knowledge bases is able to discover relevant concepts beyond the first twenty words from the topic signatures without increasing the density of the final knowledge bases. Additionally, reordering seems to provides better results than just applying a disambiguation algorithm. However, it also seems that SSI-Dijkstra (used for building KnowNet-20) outperforms *personalized PageRank* when identifying the senses of the topic signatures since KnowNet-20 outperforms deepKnowNetFirst20tswebRank. We should recall that contrary to SSI-Dijkstra, *personalized PageRank* do not use the topic signature as context.

4. Conclusions and future work

This paper reports a preliminary study exploring a new method for building KnowNets, named deepKnowNets.
Basically, the new method explores the first hundred words of every topic signature, and selects the most relevant concepts according to a similarity measure provided by UKB (Agirre and Soroa, 2009) on an *initial* knowledge base consisting on WordNet (Fellbaum, 1998) and eXtended WordNet (Mihalcea and Moldovan, 2001).
In order to evaluate the new deepKnowNets, we carried out two tasks. Firstly, we studied the different reordering of words in the original and the new topic signatures. Secondly, we compared the accuracy of the previous KnowNets with respect the new deepKnowNets in the same similarity task.
Firstly, the study on the reordering shows that the new rankings generated using the similarity measure provided by UKB are very different from the original ones. Secondly, the evaluation on the similarity task, based on human judgements annotations, denotes that deepKnowNets clearly outperform previous versions of KnowNets, and those knowledge bases required for their construction. Finally, it also seems that this new method for building automatically knowledge bases is able to discover relevant concepts beyond the first twenty words from the topic signatures without increasing the density of the final knowledge bases.

Acknowledgments

This work has been possible thanks to the support of the FP7 PATHS project (FP7-ICT-2010-270082) and the Spanish national project KNOW2 (TIN2009-14715-C04-03/04.)

5. References

Eneko Agirre and Oier Lopez de Lacalle. 2004. Publicly available topic signatures for all wordnet nominal senses. In *Proceedings of the 4rd International Conference on Language Resources and Evaluations (LREC). Lisbon, Portugal, pp. 1123-1126. ISBN: 2 - 9517408 - 1 - 6.*

Eneko Agirre and David Martínez. 2002. Integrating selectional preferences in wordnet. In *Proceedings of the 1st International Conference of Global WordNet Association*, Mysore, India.

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Eneko Agirre, Olatz Ansa, David Martínez, and Edward Hovy. 2001. Enriching wordnet concepts with topic signatures. In *Proceedings of the NAACL worshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburg, USA.

Jordi Atserias, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Piek Vossen, and Bernardo Magnini. 2004. The meaning multilingual central respository. In *Proceedings of the Second International Global WordNet Conference (GWC'04).*

Montse Cuadros and German Rigau. 2006. Quality assessment of large scale knowledge resources. In *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, Sidney, Australia, July.

Montse Cuadros and German Rigau. 2008. Knownet: using topic signatures acquired from the web for building automatically highly dense knowledge bases. In *Proceedings of 22nd International Conference on Computational Linguistics (COLING'08)*, Manchester, UK, August.

Christiane Fellbaum. 1998. *WordNet. An Electronic Lexical Database*. Language, Speech, and Communication. The MIT Press.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Claudia Leacock, Martin Chodorow, and George Miller Miller. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–166.

Chin-Yew Lin and Edward Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING*, Saarbrücken, Germany.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Dekang Lin and Patrick Pantel. 2002. Concept discovery from text. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bernardo Magnini and G. Cavaglià. 2000. Integrating subject field codes into wordnet. In *Proceedings of the Second Internatgional Conference on Language Resources and Evaluation (LREC)*, Athens. Greece.

Diana McCarthy, Rob Koeling, Julie Weeds, and John A. Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the Association for Computational Linguistics (ACL'04)*, pages 279–286, Barcelona, Spain.

Rada Mihalcea and Dan Moldovan. 2001. extended wordnet: Progress report. In *Proceedings of NAACL Workshop* WordNet and Other Lexical Resources: Applications, Extensions and Customizations, pages 95–100, Pittsburg, PA, USA.

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

Dan Moldovan and R. Gîrju. 2000. Domain-specific knowledge acquisition and classification using wordnet. In *Proceedings of FLAIRS-2000 conference*, Orlando, Fl.

I. Niles and Adam Pease. 2001. Towards a standard upper ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, Maine.

Patrick Pantel. 2005. Inducing ontological co-occurrence vectors. In *Association for Computational Linguistics (ACL-05). pp. 125-132. Ann Arbor, MI.*

Philip Resnik. 1993. *Selectional Restrictions and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proc. of IJCAI*, 14:448–453.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING/ACL*, Sidney, Australia, July.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of 16th international World Wide Web conference (WWW 2007)*, New York, NY, USA. ACM Press.

Piek Vossen. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.

Javier Àlvez, Jordi Atserias, Jordi Carrera, Salvador Climent, Egoitz Laparra, Antoni Oliver, and German Rigau. 2008. Complete and consistent annotation of wordnet using the top concept ontology. In *Proceedings of the the 6th Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech (Morocco), May.