Word Sense Disambiguation using Aggregated Similarity based on WordNet Graph Representation

Mădălina ZURINI
Academy of Economic Studies, Bucharest, Romania
madalina.zurini@gmail.com

The term of word sense disambiguation, WSD, is introduced in the context of text document processing. A knowledge based approach is conducted using WordNet lexical ontology, describing its structure and components used for the process of identification of context related senses of each polysemy words. The principal distance measures using the graph associated to WordNet are presented, analyzing their advantages and disadvantages. A general model for aggregation of distances and probabilities is proposed and implemented in an application in order to detect the context senses of each word. For the non-existing words from WordNet, a similarity measure is used based on probabilities of co-occurrences. The module of WSD is proposed for integration in the step of processing documents such as supervised and unsupervised classification in order to maximize the correctness of the classification. Future work is related to the implementation of different domain oriented ontologies.

***Keywords**: WSD, Similarity Measure, WordNet, Ontology, Synset*

1 Introduction

For the acquisition of knowledge in artificial intelligence, two approaches defined in [1] are used:

- *transfer process between human to knowledge base,* process with a major disadvantage given by the fact that the one who has knowledge cannot easily identify it;
- *conceptual modeling process* by building models in which are placed the new knowledge as they are acquired, this process leading to the appearance of the ontology as a systematic organization of knowledge, data of the reality, leading to the construction of theories upon what it exists.

An essential role of ontology is to be reused in multiple applications. Mapping two or more ontologies is called alignment. This task is particularly difficult, the main cause of limitation in extending existing ontologies [1].

Direction that follows the ontology is supported by the introduction of artificial intelligence techniques to emulate the mental representation of concepts used, and the interpenetration of these links.

The kernel of the ontology is defined as a system $\mathcal{O} = (\mathcal{L}, \mathcal{F}, \mathcal{C}^*, \mathcal{H}, ROOT)$, where:

- $\mathcal{L}$ is the lexicon formed out of the terms from the natural language;
- $\mathcal{C}^*$ a set of concepts;
- $\mathcal{F}$ represents the reference function that maps the set of terms of the lexicon to the set of concepts;
- $\mathcal{H}$ is the hierarchy of the taxonomy given by the direct, acyclic, transitive and reflexive relation;
- $ROOT$ is the starting point upon which the hierarchy is built on.

There are two types of ontologies as defined in [1], depending on the area in which they are used:

- ontologies for knowledge-based systems are characterized by a relatively small number of concepts, but linked by a large and varied relationships, concepts are grouped into complex conceptual schemes or scenarios and for each concept there can be one or more customizations;
- lexicalized ontologies, including a large number of concepts linked by a small number of relationships, like WordNet ontology concepts that are represented by sets of synonymous words, these

DOI: 10.12948/issn14531305/17.3.2013.15

ontologies are used in human language processing systems.

It is introduced the concept of ontology as a knowledge base in the classification of documents, in order to analyze semantic documents by solving the ambiguity of the terms.

This integration results in an improvement in the objective function defined for classification techniques used. The main components of an ontology are described, the concepts and relations between them. These components are analyzed, identifying methods of extracting knowledge from within.

With the defined relationships between concepts it is created the graph representation seen as a taxonomy of belonging such as "is-a" of the concepts to the more general ones. The senses of a concept are defined, along with the possibility of graph representation of each sense. In the context of WordNet ontology, the concept of synset is introduced as an equality relation between concepts with similar senses. The graph representation is further used for evaluating the similarity between two concepts. The more similar the concepts, the less the length of the path between the two nodes related to the elements in the graph representation. Two elements from the same synset maximize the similarity measure. Similarity calculation is used in the evaluation of context senses of polysemy words, measuring the maximum probability of occurrence of each sense of each words from a phrase.

2 Components and Structure of WordNet Lexical Ontology

WordNet is a database that contains information about English vocabulary. Originally designed as a full-scale model of semantic organization, was soon accepted in natural language processing NLP, Natural Language Processing. WordNet ontology has become the chosen database NLP, Kilgariff saying that not using this resource requires explanation and justification, [19]. Ontology popularity is high due to open access and wide area coverage.

WordNet ontology is created and maintained by Princeton University, the database can be downloaded from [2]. It contains nouns, verbs, adjectives and adverbs. Lexical meanings are relations between them. Words with similar meanings are organized into sets called synsets. The latest version of WordNet 3.0 contains about 155,000 words organized in 117,000 synsets, [3]. A similar synset consists of words that end with a definition and examples of use of these words.

Table 1 contains a statistic of the number of synsets existing along with the type of words from which they are formed.

Table 1. WordNet statistics, [4]

Word category	Number of unique words	Number of synsets	Total number of word-pair of senses
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Total	155287	117659	206941

Table 2 represents statistics regarding the mean number of senses of a word in the WordNet ontology.

DOI: 10.12948/issn14531305/17.3.2013.15

Table 2. Polysemy statistics [4]

Word category	Words with one sense	Polysemy words	Number of senses	Mean number of senses	Mean number of senses for polysemy words
Noun	101863	15935	44449	1.24	2.79
Verb	6277	5252	18770	2.17	3.57
Adjective	16503	4976	14399	1.40	2.71
Adverb	3748	733	1832	1.25	2.50
Total	128391	26896	79450	-	-

Areas of WordNet ontology is a lexical resource in which synsets are semi-automatic marked with one or more classes of membership in a set consisting of 165 hierarchically organized domains [5].

WordNet ontology is integrated into the representation and processing of documents as a component that solves problems like [6]:

- ignorance of any relationship between words;
- high dimensionality of the space of representation.

In [7], WordNet structure is seen as intuitive, consisting of words that have multiple meanings, each sense forming a synsets, WordNet ontology atomic structure, and relationships between words, such as synonyms, antonyms, links represented by a graph.

3 Graph representation of WordNet components

In the WordNet ontology, there are defined types of semantic relations between concepts represented by words and multiple meanings of words. Table 3 shows examples of the six types of relationships existing in the case of nouns.

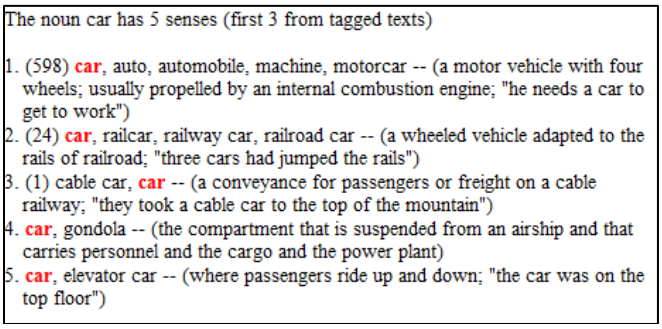
Table 3. WordNet semantic relations

Semantic relation	Syntax category	Examples
Synonymy	N, V, Aj, Av	rise, ascend
Antonymy	N, V, Aj, Av	wet, dry
Hyponymy	N	Maple, tree
Hypernymy	N	Tree, Maple
Meronymy	N	Gin, martini
Holonymy	N	Martini, Gin
Note: N- Noun V- Verb Aj- Adjective Av- Adverb		

Tree representation of the links between concepts is based on the WordNet ontology tree creating a form of words/synsets represented by nodes and links, arcs, represented by types of WordNet semantic relations between concepts. Top-bottom representation consists of a root, the point at which splits all existing links between concepts, which is called the root entity.

For the concept *car* in the WordNet ontology there are five ways identified with description and structure to the existing synset for each sense individually, Figure 1, using WordNet 2.1 Browser.

DOI: 10.12948/issn14531305/17.3.2013.15


The noun car has 5 senses (first 3 from tagged texts)

1. (598) **car**, auto, automobile, machine, motorcar -- (a motor vehicle with four wheels; usually propelled by an internal combustion engine; "he needs a car to get to work")
2. (24) **car**, railcar, railway car, railroad car -- (a wheeled vehicle adapted to the rails of railroad; "three cars had jumped the rails")
3. (1) cable car, **car** -- (a conveyance for passengers or freight on a cable railway; "they took a cable car to the top of the mountain")
4. **car**, gondola -- (the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant)
5. **car**, elevator car -- (where passengers ride up and down; "the car was on the top floor")

Fig. 1. Senses of car noun from WordNet ontology

Each sense becomes leaf node for the semantic graph representation using semantic relations. Figure 2 generated using WordNet Browser 2.1 contains an example of a graph representing the first sense of the concept *car* using WordNet semantic relations in the ontology.

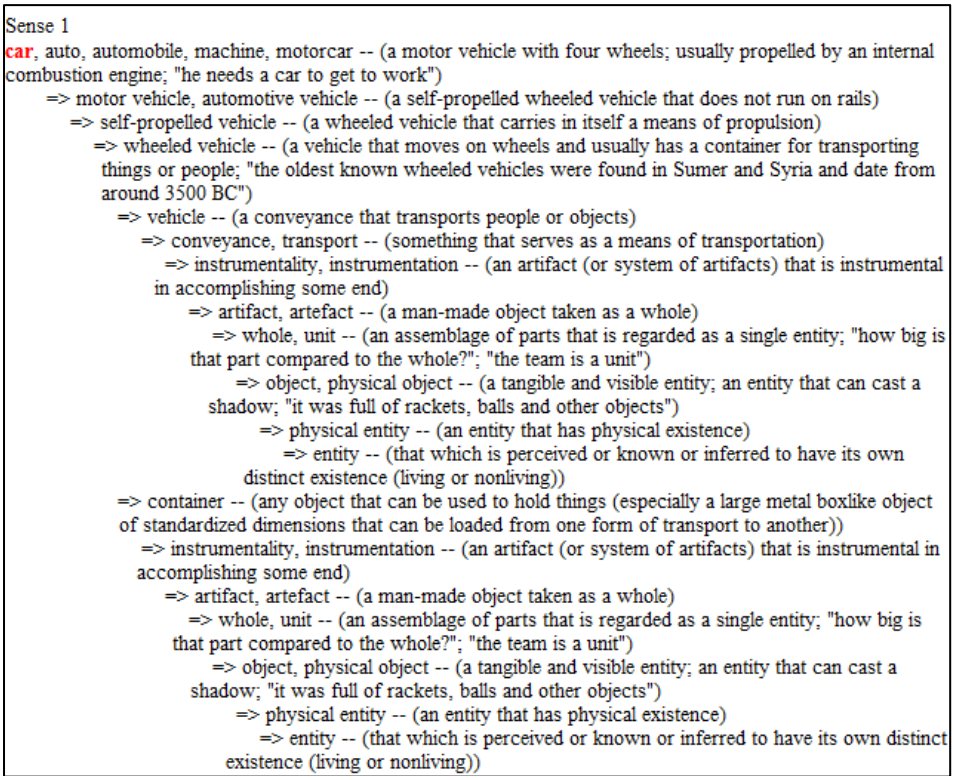

Sense 1
car, auto, automobile, machine, motorcar -- (a motor vehicle with four wheels; usually propelled by an internal combustion engine; "he needs a car to get to work")
=> motor vehicle, automotive vehicle -- (a self-propelled wheeled vehicle that does not run on rails)
=> self-propelled vehicle -- (a wheeled vehicle that carries in itself a means of propulsion)
=> wheeled vehicle -- (a vehicle that moves on wheels and usually has a container for transporting things or people; "the oldest known wheeled vehicles were found in Sumer and Syria and date from around 3500 BC")
=> vehicle -- (a conveyance that transports people or objects)
=> conveyance, transport -- (something that serves as a means of transportation)
=> instrumentality, instrumentation -- (an artifact (or system of artifacts) that is instrumental in accomplishing some end)
=> artifact, artefact -- (a man-made object taken as a whole)
=> whole, unit -- (an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit")
=> object, physical object -- (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")
=> physical entity -- (an entity that has physical existence)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
=> container -- (any object that can be used to hold things (especially a large metal boxlike object of standardized dimensions that can be loaded from one form of transport to another))
=> instrumentality, instrumentation -- (an artifact (or system of artifacts) that is instrumental in accomplishing some end)
=> artifact, artefact -- (a man-made object taken as a whole)
=> whole, unit -- (an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit")
=> object, physical object -- (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")
=> physical entity -- (an entity that has physical existence)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Fig. 2. Is-a relations for the first sense of car noun

Based on the relations of *"is-a"* type, the graph representation is formed, Figure 3.

DOI: 10.12948/issn14531305/17.3.2013.15

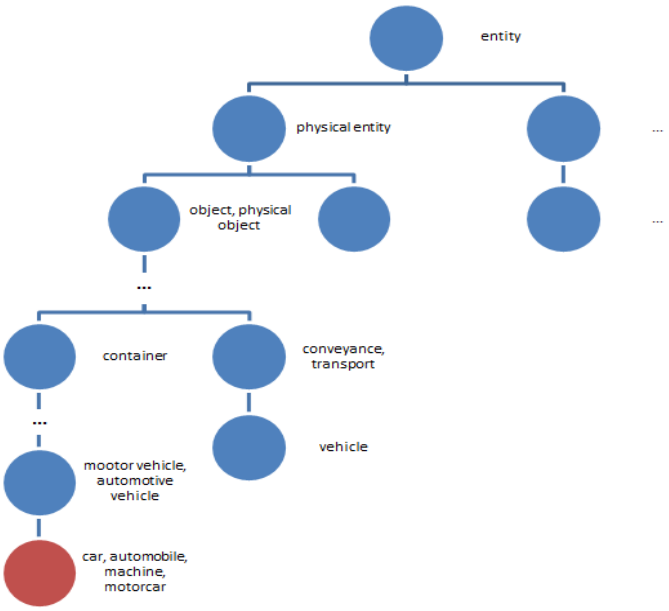


Fig. 3. Graph associated to the first sense of car noun using is-a relations

This metric is then used in the evaluation of applications for text documents, as well as supervised classification and clustering, the semantic problem solving.

4 Similarity Measure of Strength Connection between Two Nouns

The similarity between the two concepts in the is-a hierarchy of the graph associated to the ontology WordNet quantifies how much resemble those objects based on information held on schedule [8]. Measurement correlation and the distance between words is used in applications such as identifying contextual meanings of words, determining the structure of text documents, creating automatic summaries, information extraction and automatic indexing [9].

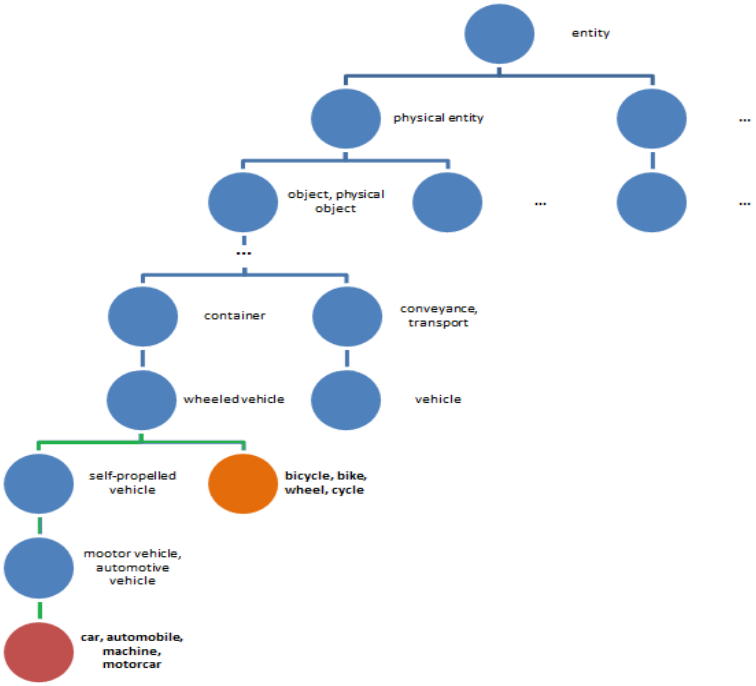


Fig. 4. Graph associated to car and bicycle nouns using is-a relations

DOI: 10.12948/issn14531305/17.3.2013.15

For understating the way of similarity calculation between the WordNet concepts, the graph associated to the WordNet ontology is given as starting point. Figure 4 contains part of the representation for the examples car and bicycle.

In the context of similarity identification between c_1 and c_2 concepts with multiple senses, the metric result is given by the maximum between the values of the similarity metric of each senses of concepts c_1 and c_2.

For that, noting the general similarity measure with $d_{SIM}: C \times C \longrightarrow \mathbb{R}^+$, where C represents the set of concepts existed in the graph G, the similarity value is given by the relation, [10]:

$$d_{SIM}(c_1, c_2) = \max_{c_{1\#} \in sens(c_1), c_{2\#} \in sens(c_2)} d_{SIM}(c_{1\#}, c_{2\#})$$

where:

- $sens(c)$ represents the set of senses of the concept c, with $c \in G$;
- $c_{\#}$ represents a sense from the set of senses associated for concept c.

Table 4 contains the formulas for measuring the correlation or similarity between two concepts from the WordNet ontology, [11], [LINGLI12] and [12]. Two categories of similarity measures exist, [10] and [12]:

- based on edge relation from which the graph G is formed of;
- based on the information retained in the nodes of the WordNet graph by adding analysis upon priory existing set of documents.

The general model of abstracting of similarity measure, [13], based on the edge relation is given by the formula:

$$d_{SIM}(c_1, c_2) = 2 \cdot \gamma - (\alpha + \beta)$$

where α, β, γ represents the attributes of concepts c_1, c_2 and their closest parent.

Table 4. Formulas for metrics for evaluation of similarity between two concepts of WordNet

Correlation metric	Calculation formula	Variables used
Path Length	$d_{PATH}(c_1, c_2) = \frac{1}{lg(c_1, c_2)}$	$lg(c_1, c_2)$ the minimum length between c_1 and c_2 nodes.
Leacock & Chodorow	$d_{LC}(c_1, c_2) = -\log \frac{lg(c_1, c_2)}{2 \cdot \max_{c \epsilon G} \lg(c)}$	G graph associated to WordNet ontology.
Wu & Palmer	$d_{WP}(c_1, c_2) = \frac{2 \cdot lg(l(c_1, c_2))}{lg(c_1, l(c_1, c_2)) + lg(c_2, l(c_1, c_2)) + lg(c_1, l(c_1, c_2))}$	$l(c_1, c_2)$ first mutual parent of concepts c_1 and c_2.
Resnik	$d_{RE}(c_1, c_2) = -\log p(l(c_1, c_2))$	$p(c)$ probability of occurrence of concept c.
Jiang & Conrath	$d_{JC}(c_1, c_2) = 2 \cdot \log p(l(c_1, c_2)) - (\log p(c_1) + \log p(c_2))$	
Lin	$d_{LIN}(c_1, c_2) = \frac{2 \cdot \log p(l(c_1, c_2))}{\log p(c_1) + \log p(c_2)}$	

Three similarity metrics are based upon the path between two concepts c_1 and c_2: Leacock & Chodorow, Wu & Palmer and Path Length. The metric $d_{LC}(c_1, c_2)$ identifies the shortest path between c_1 and c_2 and scales this value to the maximum path length from the *is-a* hierarchy in which they appear, [8]. The metric $d_{WP}(c_1, c_2)$ identifies the length of the path form the root node to the closest mutual parent of the two elements, which is normalized by the amount of each individual object to the root. Metric $d_{PATH}(c_1, c_2)$ measures the reverse path length between c_1 and c_2.

The metrics $d_{JC}(c_1, c_2)$, $d_{RE}(c_1, c_2)$ and $d_{LIN}(c_1, c_2)$ are based on the specificity of the analyzed concepts. $d_{LIN}(c_1, c_2)$ and $d_{JC}(c_1, c_2)$ increases the information retained

DOI: 10.12948/issn14531305/17.3.2013.15

of the closest node mutual parent with the sum of information of each individual concept. The information retained by each node is derivative from the label of the senses of the SemCor set of documents. Different approaches also exist using Brown Corpus, the Penn Treebank or the British National Corpus.

Based on the relations between the two concepts car and bicycle, table 5 contains the similarity metric values defined in Table 4.

Table 5. Similarity metrics' values between car and bicycle nouns

Correlation metric	**Word 1**	**Word 2**	**Score**
Path Length	Car#s2	Bicycle#s1	0.33
	Car#s1	Bicycle#s1	0.2
	Car#s3	Bicycle#s1	0.1
	Car#s4	Bicycle#s1	0.1
	Car#s5	Bicycle#s1	0.1
Leacock & Chodorow	Car#s2	Bicycle#s1	2.59
	Car#s1	Bicycle#s1	2.07
	Car#s3	Bicycle#s1	1.38
	Car#s4	Bicycle#s1	1.38
	Car#s5	Bicycle#s1	1.38
Wu & Palmer	Car#s2	Bicycle#s1	0.9
	Car#s1	Bicycle#s1	0.81
	Car#s3	Bicycle#s1	0.57
	Car#s4	Bicycle#s1	0.57
	Car#s5	Bicycle#s1	0.57
Resnik	Car#s2	Bicycle#s1	6.31
	Car#s1	Bicycle#s1	6.31
	Car#s3	Bicycle#s1	2.49
	Car#s4	Bicycle#s1	2.49
	Car#s5	Bicycle#s1	2.49
Jiang & Conrath	Car#s2	Bicycle#s1	0.22
	Car#s1	Bicycle#s1	0.14
	Car#s3	Bicycle#s1	0
	Car#s4	Bicycle#s1	0
	Car#s5	Bicycle#s1	0
Lin	Car#s1	Bicycle#s1	0.73
	Car#s2	Bicycle#s1	0.64
	Car#s3	Bicycle#s1	0
	Car#s4	Bicycle#s1	0
	Car#s5	Bicycle#s1	0

Regardless of the correlation metric used, the second meaning of the noun *car* is in maximum correlation to *bicycle* noun, a lower score being obtained only for the metric developed by Lin.

In [14], an analysis is made of the metric to calculate the similarity by identifying the minimum path length between two concepts represented as nodes in the graph associated ontology WordNet.

Applying the metric of the minimum length between two nodes is a correct measure of semantic distance in the case where the density of the terms across the semantic network is constant. But how general semantic network density is not constant, the number of nodes in the network increases with deepening in direct correlation with the increasing number of terms is required densities approach along with the shortest path evaluation.

An example that reinforces this idea is given by the differences between the sets of concepts *{plant, animal}* and *{zebra, horse}*,

DOI: 10.12948/issn14531305/17.3.2013.15

sets of concepts of a 2-link both, but the connection between the first two concepts is lower than the next two. This difference is given by the position at which the concepts are situated from the root level. *Plant* and *animal* concepts are more general, situated at a superior level, beside *zebra* and *horse* concepts, more particular ones.

By applying the simple process of calculating the depth of a node, the shortest path length metric is significantly improved. Problem which is reached is the transposition of the depth of a node into a density.

The work of Richardson [15] suggests using the value of the density the depth calculated of each node itself. Thus, the distance between two nodes is calculated as the ratio between the length and the density of the minimum distance between nodes of the graph. As this method involves a linear relationship between depth and density, an assumption is not true in all cases; it is proposed to calculate the average density for each level of the graph. It is created a function associating a graphical averaged density. Let *FDM* be the function that receives as a parameter the graph level and returns the average density of that level. Using this approach, the distance function between two nodes is:

$$d_{DW_{PATH}}(x,y) = \frac{d(x,y)}{FDM(L)}$$

where:

- $d_{DW_{PATH}}$ is the weighted distance between two WordNet concepts;
- $d(x,y)$ is the minimum length of the path between *x* and *y*;
- *L* is the level where *x* and *y* nodes are found within the graph.

Since two nodes not necessarily are found at the same level, a way of solving this problem consists in assigning the level L with the level where the closest parent of *x* and *y* nodes is part of.

5 Word Sense Disambiguation of Polysemantic Nouns

Automatic evaluation of contextual meanings of words had an interest and concern since the beginning of natural language processing. Evaluation meaning is not seen as an independent business, but as an intermediate step and necessary in order to achieve the semantic processing of text objects, [16].

One way to solve the problem of choosing the contextual meaning of a word in the context in which the word is polysemy is to extend analysis at word level way, increasing the size of the representation of text documents directly proportional to the number of senses added in the analysis, and training base to be able to perform statistical analysis of the occurrence of contextual meanings, and correlations with other words that deal directly.

The base from which to start analyzing the contextual meaning of a word is the number of meanings available in WordNet ontology, along with a counter of the number of times meaning emerged. Figure 5 contains meanings and word counting *car*.

The noun car has 5 senses (first 3 from tagged texts)

1. (598) **car**, auto, automobile, machine, motorcar -- (a motor vehicle with four wheels; usually propelled by an internal combustion engine; "he needs a car to get to work")
2. (24) **car**, railcar, railway car, railroad car -- (a wheeled vehicle adapted to the rails of railroad; "three cars had jumped the rails")
3. (1) cable car, **car** -- (a conveyance for passengers or freight on a cable railway; "they took a cable car to the top of the mountain")
4. **car**, gondola -- (the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant)
5. **car**, elevator car -- (where passengers ride up and down; "the car was on the top floor")

Fig. 5. Senses and number of appearance of car noun

DOI: 10.12948/issn14531305/17.3.2013.15

Table 6 contains the probabilities of occurrence of the word car senses calculated as part from the whole, reported to the number of appearance of the analyzed sense to the total number of appearances of the noun car, regarding the contextual sense.

Table 6. Appearance probabilities of car's senses

Sense	Probability
car#s1	$P(car\#s1) = \frac{598}{623} = 95.98\%$
car#s2	$P(car\#s2) = \frac{24}{623} = 3.85\%$
car#s3	$P(car\#s3) = \frac{1}{623} = 0.16\%$
car#s4	$P(car\#s4) = \frac{0}{623} = 0\%$
car#s5	$P(car\#s5) = \frac{0}{623} = 0\%$

If the sense of the noun car is chosen to by the first, regardless of its contextual analysis, using the statistics offered within WordNet ontology, the percentage of error of assignation would be 4.02%. This situation, instead, isn't as favorable for each concept from the WordNet ontology. For that, an appearance evaluation of each word is needed. Different studies focus on identifying the methods of evaluating the contextual senses of polysemy words, such as in [17].

The kernel of the sense disambiguation algorithm consists in computing the semantic similarity using the taxonomy of WordNet ontology, [18].

The general model for describing the problem of context sense choosing is given by the existence of a set of key words from which a phrase is formed of:

$$F = \{w_1, w_2, \dots, w_f\}$$

where:

- F is the analyzed phrase;
- f is the number of words;
- w_i is the i word of the phrase F.

For each word from F phrase, the set of senses is formed:

$$s_i = \{s_{i1}, s_{i2}, \dots, s_{Card(s_i)}\}$$

where s_{ij} is the j sense of word w_i.

Let w_p be a polysemy word, the problem of sense identification summarizes in choosing the sense that has maximum similarity between the word and each other word found in the phrase. Using the similarity measures previously described, the maximization model is:

$$w_p = \left\{ s_{pj} \middle| j = arg \max_{j=1,Card(s_p)} \frac{\sum_{\substack{k=1 \\ k \neq p}}^{f} d_{SIM}(w_{pj}, w_k)}{f-1} \right\}$$

If more polysemy words exist, the algorithm is repeated for each one, resulting in a sense that maximizes the context similarity. Each polysemy word is analyzed according to the words and senses found after it. Comparing the probabilities for the first polysemy word with the rest, the resulted sense maximizes the semantic information.

Let F_1 and F_2 be the two phrases formed out of the key words:

$$F_1 = \{road, car, speed\}$$
$$F_2 = \{road, success, fame\}$$

The word road has two senses available in WordNet ontology, described in Figure 6.

The noun road has 2 senses (first 2 from tagged texts)

1. (95) **road**, route -- (an open way (generally public) for travel or transportation)
2. (2) **road** -- (a way or means to achieve something; "the road to fame")

Fig. 6. Senses of word road

DOI: 10.12948/issn14531305/17.3.2013.15

For evaluating the senses of word road for each phrase F_1 and F_2 the array of similarities using the d_{PATH} metric between road word and the rest existing words is formed, Table 7.

Table 7. Similarity values between the words from F_1 and F_2

Similarity	car	speed	success	fame
road#s1	0.1429	0.0910	0.1250	0.0714
road#s2	0.0625	0.1250	0.1429	0.0769

Table 8 contains the aggregated probabilities for the two senses of word road according to each phase.

Table 8. Probabilities of the senses of word road

Probability	F_1	F_2
road#s1	0.1169	0.0982
road#s2	0.0937	0.1099

For the first phrase, the sense chosen for the word road is the first one, and for the second phrase, the sense of the word road is *road#s2*.

Once selected the sense of road word, the next polysemy word is analyzed, car. An optimization method consists in choosing only the senses upon which there are statistics in WordNet ontology, for the noun car mentioning the 1, 2 and 3 senses, table 9.

Table 9. Similarity measures of car senses and the other words

Similarity	road#s1	speed
car#s1	0.1111	0.0667
car#s2	0.1429	0.0769
car#s3	0.1250	0.0667

Table 10 contains the aggregated probabilities for determining the dominant sense that maximizes the semantic similarity.

Table 10. The probabilities of noun car

Probability	F_1
car#s1	0.0888
car#s2	0.1099
car#s3	0.0958

Because of the fact that the aggregated probability for the second sense of the noun car is greater, the contextual sense of the word car is *car#s2*.

In Figure 7, it is presented the source code of the WSD process.

```
txt_results.Visible = true;
txt_results.Text = "";
words = txt_words.Text.ToString().Split(' ');

WNCommon.path = @"C:\Program Files\WordNet\2.1\dict\";

MyWordInfo[] words_info = new MyWordInfo[words.Length];
for (int i = 0; i < words_info.Length; i++)
{
    words_info[i] = new MyWordInfo(words[i], Wnlib.PartsOfSpeech.Noun);
}
WordSenseDisambiguator WSD = new WordSenseDisambiguator();
MyWordInfo[] info = WSD.Disambiguate(words_info);
```

Fig. 7. Source code for WSD process

The testing process consists in running a set of phrases priory contextual sense classified. The metric used for evaluating the WSD correctness, IC_{WSD}, is defined using:

$$IC_{WSD} = \frac{\sum_{i=1}^{nr\_wsd} w_i}{nr\_wsd} \times 100$$

where:

DOI: 10.12948/issn14531305/17.3.2013.15

- *nr_wsd* is the number of polysemy words existing in the phrases used for testing;
- w_i represents the association between the priory sense of the i word with the sense generated by WSD algorithm, based on the formula:

$$w_i = \begin{cases} 1, if\ sens\_aprioric_i = sensWSD_i \\ 0, otherwise \end{cases}$$

- $sens\_aprioric_i$ is the priory sense associated to the word *i*;
- $sensWSD_i$ is the sense generated by WSD algorithm for the *i* word.

A testing set formed out of 100 phrases is used, containing *nr_wsd=200* polysemy words. After running WSD algorithm, the value of IC_{WSD} indicator is 94%.

6 Conclusions

Adding a context analysis for the words that has multiple meaning according to the neighbor words increases the performance of text document processing and representation. The proposed aggregation method of the probabilities of each sense of the existing words within a phrase optimizes the correctness of the word sense disambiguation process, taking into account the space and time consuming elements.

WordNet ontology is added as an external knowledge base used for an up level representation of English concepts, resolving the problem of similarity among the existing concepts.

The results of the WSD process indicate a correctness level of 94% for the testing set used.

Acknowledgments

This work was cofinaced from the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/107/1.5/S/77213 „Ph.D. for a career in interdisciplinary economic research at the European standards".

References

[1] S. Trausan-Matu, *Inteligenta artificiala*, 2004, Available online at : http://www.racai.ro/~trausan/ia.pdf

[2] WordNet. *A lexical database for English*, Available online at: http://wordnet.princeton.edu/wordnet/related-projects/

[3] E. Hessami, F. Mahmoudi, H. Jadidinejad, "Unsupervised Graph-based Word Sense Disambiguation Using lexical relation of WordNet", *International Journal of Computer Science Issues*, Vol. 8, Nr. 3, 2011, pg. 225-230

[4] WordNet Statistics: Available online at: http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

[5] A. Gonzalez, G. Rigau, M. Castillo, "A graph-based method to improve WordNet Domains," *Proceeding CICLing'12 Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing*, Vol. 1, 2012, pg. 17-28.

[6] Z. Elberrichi, A. Rahmoun, M. A. Bentaalah, "Using WordNet for Text Categorization", *The International Arab Journal of Information Technology*, Vol. 5, Nr. 1, 2008, pg. 16-24, ISSN 1683-3198

[7] A. Passos, J. Wainer, "Wordnet-based metrics do not seem to help document clustering", 2009, Available online at: http://www.ic.unicamp.br/~tachard/docs/wncluster.pdf

[8] T. Pedersen, S. Patwardhan, J. Michelizzi, "WordNet::Similarity – Measuring the Relatedness of Concepts", *Proceeding HLT-NAACL--Demonstrations '04 Demonstration Papers at HLT-NAACL*, May, 2004, Boston, pg. 38-41

[9] A. Budanitsky, G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness", *Journal Computational Linguistics*, Vol. 32. Nr. 1, 2006, pg. 13-47.

[10] Q. Peng, L. Zhao, Y. Yu, W. Fang, "A New Measure of Word Semantic Similarity based on WordNet Hierarchy and DAG Theory," *International Conference on Web Information Systems*

DOI: 10.12948/issn14531305/17.3.2013.15

and Mining, 2009, pg. 181-185, ISBN 978-0-7695-3817-4

[11] E. Blanchard, M. Harzallah, H. Briand, P. Kuntz, "A typology of ontology-based semantic measures", *Proceeding of EMOI-INTEROP 05*, Portugal, June 2005

[12] A. Buhanitzky, G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness", *Journal Computational Linguistics*, Vol. 32, Nr. 1, 2006, pg. 13-47, ISSN 1530-9312

[13] F. Lin, K. Sandkuhl, "A Survey of Exploiting WordNet in Ontology Matching", *IFIP International Federation for Information Processing*, Vol. 276, Artificial Intelligence and Practice II, Boston, Springer, 2008, pg. 341-350

[14] D. Yang, D. M. W. Powers, "Measuring Semantic Similarity in the Taxonomy of WordNet", *28th Australasian Computer Science Conference*, Newcastle, Australia, 2005, pg. 315-322

[15] W. D. Lewis, "Measuring Conceptual Distance Using WordNet: The Design of a Metric for Measuring Semantic Similarity", *Language in Cognitive Science*, 2001, pg. 9-16, Available online at: http://coyotepapers.sbs.arizona.edu/CPXII/Lewis.pdf

[16] R. Richardson, A. Smeaton, J. Murphy, "Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Word", *Technical Report, Working paper CA-1294*, School of Computer Applications, Dublin City University, 1994

[17] S. Kamali, "Some Experiments in Word Sense Disambiguation", 2001, Available online at: https://cs.uwaterloo.ca/~s3kamali/courses/word-sense-disambiguation.pdf

[18] L. Xiaobin, S. Szpakowicz, S. Matwin, "A WordNet-based Algorithm for Word Sense Disambiguation", *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pg. 1368—1374

[19] P. Resnik, "Disambiguating Noun Grouping with Respect to WordNet Senses," *Natural Language Processing Using Very Large Corpora Text, Speech and Language Technology*, Vol. 11, 1999, pg. 77-98.

[20] J. Boyd-Graber, C. Fellbaum, D. Osherson, R. Schapire, "Adding Dense, Weighted Connections to WordNet", 2005, Available online at: https://wordnet.princeton.edu/wordnet/publications/jbj-jejufellbaum.pdf

Mădălina ZURINI is currently a PhD candidate in the field of Economic Informatics. She graduated the Faculty of Cybernetics, Statistics and Economic Informatics (2008) and a master in Computer Science, having her dissertation given in *Implications of Bayesian classifications for optimizing spam filters* (2010). She is also engaged in Pedagogical Program as part of the Department of Pedagogical Studies. Her fields of interest are data classification, artificial intelligence, data quality, algorithm analysis and optimizations. She wants to pursue a pedagogical career.

DOI: 10.12948/issn14531305/17.3.2013.15