# WordNet as a Tool for Measurement of Domain Similarity of Texts

## General statements

As we enter the era of information society, information retrieval in the Internet becomes inevitable part of our everyday life. It is universally acknowledged that the IR remains to be rather a sophisticated activity due to the large amount of irrelevant information supplied for any user's inquiry. Thus there are two challenges a researcher is facing:

1. To minimize the volume of the so called “garbage” (e.g. links that the user doesn’t need, documents irrelevant to the inquiry).
2. To bring order into the chaotic massive of texts in the Internet.

Each of the issues requires a serious research work.

The goal of the present paper is to report on the work aiming at the developing of the algorithm (and then a programme), which could sort documents (texts) from the point of view of their domains. Yet it should not be an obligatory assigning of domain labels, but a probabilistic one (e.g. a document may belong to Domain 1 with 40% probability, to Domain 2 with 33% probability; and so on up to Domain N with its, for example, 1% probability). Obviously, if the user searches documents on music he should get a list of such documents in order of the percentage in the Music Domain. Thus the amount of so-called “garbage” will be diminished.

## Experiment

We have developed an algorithm which enables us to evaluate the probability of the fact that given texts belong to the same Domain, so that related texts can be grouped together.

As a tool for similarity measurement we used Princeton WN 1.5., namely its hypernymy/hyponymy structures. The underlying idea was that generic relations between words played the crucial role in Information Retrieval (e.g., providing we've got texts about football and about tennis, it is generic relation which combines those texts into one domain "Sports")

The pilot study of the algorithm on English corpus (Financial Times collection '98) was carried out. Texts with various domains were taken: sports, politics, economics, medicine and arts.

## Procedure of analysis

In short, the following steps were made:

1. Equally sized extracts from the articles (about 2500 symbols) were taken. Each article had a preliminary Domain Label assigned manually.
2. For each significant word in text the appropriate Top Ontology unit was found. Also we took into account nodes of the next 3 lower levels in the hypernymy/hyponymy hierarchy. But they occurred to be inefficient in the course of the study. Thus although various depths of analysis were tested, only Top Ontology level provided clear positive results. The possible explanation says that the reason lays in the fact that different words have different sizes / depths / heights of the trees, and sometimes the only level they have in common is that of Top Ontology.
3. For each text we obtained the list of TO units.
4. The TO elements were weighted in accordance to special mathematic procedures. The relation between occurrence of the words in texts and the weight assigned to their most common hypernym was as follows:
    - The more frequent grandchildren (hypo-hyponyms) of the TO unit occured in the text the larger weight was assigned to it.
    - The more senses had a grandchild (hypo-hyponyms) (as they are fixed in WN) the lesser weight it assigned to its TO grandparent.

| № of sense / Number of word senses | **1** | **2** | **3** | **4** | **5** | **6** | **7…** |
|---|---|---|---|---|---|---|---|
| **1** | 1 | | | | | | |
| **2** | 0,5 | 0,5 | | | | | |
| **3** | 0,5 | 0,25 | 0,25 | | | | |
| **4** | 0,5 | 0,25 | 0,125 | 0,125 | | | |
| **5** | 0,5 | 0,25 | 0,125 | 0,063 | 0,062 | | |
| **6** | 0,5 | 0,25 | 0,125 | 0,063 | 0,031 | 0,031 | |
| **7…** | 0,5 | 0,25 | 0,125 | 0,063 | 0,031 | 0,016 | 0,015 |

**Table 1: The weights assigned to the TO units.**

**(Column 1 - total number of senses, Raw 1 - № of the sense according to the WN)**

So, if we take into consideration the text *To the thump and blare of marching bands, Belarus', independence day parade ploughed through Minsk yesterday… The event was a chance for Alexander Lukashenko, the country's stern, mustachioed president, to show off his country's march back to soviet style communism* we see that TO unit "*GROUP*" gets partial weights of 0,25 and 0,125 because of the occurrence of its hypo-hyponyms "*Band*" and "*Country*" respectively.

5. So, we could determune the absolute weight of each TO unit in the text by summirising of all its partial weights. E.g., the *"GROUP"* finally got an absolute weight of 3,719.
6. Then we calculated the average absolute TO weight for each document .
7. And finally each text was represented by the list of appropriate TO units and their weights.
8. And then for those weights lines we've calculated correlation between pairs of documents.

**Results.**

Correlation between pairs of texts belonging to the same domains has significantly exceeded that of the pairs of documents belonging to different domains (0,76 and 0,22 respectively), and thus the algorithm was considered as an efficient one.

**Problems and Discussions**

The main problem was that of Word Sense Disambiguation (WSD). We can either try to solve it manually or automatically (to achieve better results), or disregard it.
We probed different strategies of preliminary text analysis in order to estimate their validity for task of Domain Similarity Measurement:

a. In the beginning we decided to test the method in the "ideal" circumstances, i.e. that of **Manual WSD** for every significant word (when the researcher chose word sense which he considers to be right). And obviously, this strategy resulted in the best characters. But it requires too much time and efforts, so we used it just to estimate the validity of other strategies.

b. Another strategy was to take into consideration **all the WN word senses**. That is to act as a primitive programme without syntactical and morphological analysers. For example, the word "*play*" from one of the texts had been introduced by all its senses as a noun and also as a verb, amounting 17. This fact actually spoiled the results and the strategy was proved to be unsatisfactory.

c. The closest to the ideal strategy turned to be that when all the documents were **part-of-speech tagged automatically** by a syntactic analyser, which helped to reduce huge ambiguity however not to eliminate it.

Although the automatic syntactic analyser does not perform perfectly and can't eliminate ambiguity in a whole, we prefer it to manual part-of-speech tagging, which results in the best figures but in itself has significant disadvantages for our purposes.

**Perspectives**

Since the algorithm is efficient, we should test it on a bigger corpus, namely not only newspaper articles, but also fiction and scientific texts. Then some mask (image) for every Domain will be needed in order to compare with every document available. We'll also need to develop mathematic apparatus of the probabilistic prognosis able to label documents to some Domains.

**Conclusions**

The algorithm described affords us to evaluate the Domain similarity of various texts and thus to group semantically related texts together.

It is based on the WordNet hypernymy/hyponymy structure, which is core relation for all WordNets however different they are.

The main advantage of the method: it doesn't need additional assigning of the Domain labels to the words in WN**.**

As WordNets are already available for most of the European and many Asian languages, it seems reasonable to test the method for other languages. In case the results are positive, the Internet users will taste information retrieval fruits of better quality.