# WordFinder

**Catalin Mititelu**
Stefanini / 6A Dimitrie Pompei Bd,
Bucharest, Romania
catalinmititelu@yahoo.com

**Verginica Barbu Mititelu**
RACAI / 13 Calea 13 Septembrie,
Bucharest, Romania
vergi@racai.ro

## Abstract

This paper presents our relations-oriented approach to the shared task on lexical access in language production, as well as the results we obtained. We relied mainly on the semantic and lexical relations between words as they are recorded in the Princeton WordNet, although also considering co-occurrence in the Google n-gram corpus. After the end of the shared task we continued working on the system and the further adjustments (involving part of speech information and position of the candidate in the synset) and those results are presented as well.

## 1 Introduction

In this paper we present our experience in the shared task on lexical access in language production, organized as part of the CogALex workshop. Given a list of five words (let us call them seeds), the system should return a word (we will call it target) which is assumed to be the most closely associated to all the seeds. Two remarks are worth being made here: on the one hand, what we call word is in fact a word form, as inflected forms are both among the seeds and among the expected targets in the training and the test sets. On the other hand, the closeness of association remains understated by the organizers. It can be understood at several levels, given our analysis of the training data: the meaning and/or the form, the syntagmatic associations, i.e. associations of words in texts. However, our system dealt mainly with the semantic level. The form level is involved only to the extent to which lexical relations (usually derivational relations and antonymy) in Princeton WordNet (PWN) are used. The syntagmatic relations we use are the co-occurrences in the Google n-gram corpus

## 2 Our understanding of the lexical access task

Having already established what meaning we, as speakers, want to render, the lexical choice is influenced by several factors: the person we talk to, the circumstances (place, other participants) of our discussion, the social (or even other types of) relations between the participants to the discussion. The shared task focuses on the tip of the tongue (TOT) phenomenon, as rightly described in the shared task presentation: we do not remember the word "mocha", but we want to express the idea (i.e., the meaning) "superior dark coffee made of beans from Arabia". In a real life conversation, dealing with TOT is much simpler: the speaker (the one affected by TOT) has the ability of defining the word s/he is looking for or of enumerating some words AND specifying the relation(s) they establish with the looked for word. Thus, we consider that the task here, consisting of being able to find the target when receiving five seeds, does not mimic the real life situation. In fact, we deprive the system of vital information that, we, as speakers, possess, to our great advantage reflected in our success in dealing with the TOT problem, after all. Moreover, given the information provided by the organizers once the results were send, the seeds that we received are derived from the Edinburgh Associative Thesaurus, so they are, in fact, the associations introduced by the users to a seed. So, the organizers implicitly considered the association of two words is the same, irrespective of which of them is the seed and which is the target, which is definitely not the same, especially if the association is a syntagmatic one.

*Zock/Rapp/Huang (eds.): Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon*, pages 68–74,
Dublin, Ireland, August 23, 2014.

## 3 Related work

In a recent experiment (Zock and Shcwab, 2013), a set of seeds (called stimuli therein) is presented to a system and, relying on information available in the eXtended WordNet (Mihalcea, 2001) and in DBpedia, a list of words is returned. The authors explain the bad results by the small dimensions of the eXtended WordNet and by the small number of syntagmatic relations it contains. Although they emphasize the necessity of using big corpora, with heterogenous data, to help solve the TOT problem, the conclusions speculate about various elements that can lead to, but do not guarantee the success:

- the big size of the corpus, the heterogeneity of the texts it contains;
- high density of relations in a network;
- the quality of the search;
- all these together.

## 4 Our approach

### 4.1 The data

The training set contains a list of 2000 pairs of five seeds and the target. They look quite heterogeneous: there are content and functional words alike, lemmas and inflected forms (see "occurs $\sim$ happens happen often sometimes now"), capitalized (sometimes unnecessarily, for example "Nevertheless" in the pair "however $\sim$ but never Nevertheless when although") and uncapitalized words.
Interestingly, two different inflected forms are targets of (partially) different sets of seeds: compare:
occur $\sim$ happen event often perfume today
with
occurs $\sim$ happens happen often sometimes now.
This means that not only semantic relations are established between the seed and the target, but also grammatical ones.

### 4.2 Assumptions

In order to construct our system we made the assumption, supported by the manual analysis of the training set, that the seeds and the target are related to each other by different kinds of relations:

- semantic relations;
- co-occurrence, in either order;
- syntactic relations;
- gloss-like relations, i.e. the target may be defined using one or more seeds;
- domain relations, i.e. the target and at least some seeds may belong to the same domain;
- form relation, i.e. the target and one or more seeds may display a partial identity of form (and sometimes even of the acoustic form of words);
- inflection as a relation among forms of the same word;
- etc.

Given these, we were aware of the impossibility of dealing with cases involving inflected forms, some of them occurring as seeds, while one occurs as target, such as:
am $\sim$ I not is me are.
In this case, an inflectional relation can be found between "is" and "am" and between "are" and "am", whereas the relations between "am" and "I" and between "am" and "not" are syntagmatic (co-occurrences). No relation can we identify between "am" and "me".

### 4.3 Resources

As a consequence of the assumptions made, the language resources we used for the competition were the Princeton WordNet (PWN) (Fellbaum, 1998) and Google n-grams corpus (Brants and Franz, 2006). The implied limitations of our approach are:

- the impossibility of dealing with pairs involving only inflected words (as in the previous example) or only functional words (as in the case: "at ∼ home by here in on");
- no contribution made by some of the seeds in the process of finding the target;
- the partial dealing with inflected forms such as plurals, third person singular of verbs, gerunds, as they cannot be found in PWN; the only source of information about them is the n-grams corpus;
- some combinations (although quite frequent, according to our intuitions obout the language) cannot be found in the Google n-gram corpus.

For all (2000x5) pairs seed-target in the training set we extracted from PWN the shortest relations chains, as a kind of lexical chains (Moldovan and Novischi, 2002), existing between them, disregarding the part of speech of the words. These chains are made up of both semantic and lexical relations (as they are defined in the wordnet literature, i.e. lexical relations are established between word forms, while semantic relations are established between word meanings). The most frequent relations chains are presented in Table 1. Straightforwardly, the most frequent association between the seeds and the targets (occurring 548 times) is of the kind synonymy. However, various combinations of hyponymy and hypernymy account for a significant number of pairs: 1213. Almost half of these cases (510) are solved by only one of the two relations (328 by hyponymy alone and 182 by hypernymy alone). Moreover, these relations contribute also in chains involving the derivat relation. So, we can consider them the most useful ones. (Our finding is similar to the weight associated to these relations by Moldovan and Novischi (Moldovan and

| **Lexical chain** | **Number of occurrences** |
|---|---|
| synonym | 548 |
| hypernym hyponym | 332 |
| hyponym | 328 |
| hypernym | 182 |
| antonym | 143 |
| similar_to | 128 |
| derivat | 119 |
| hypernym hyponym hyponym | 115 |
| hypernym hypernym hyponym | 100 |
| hyponym hyponym | 81 |
| hypernym hypernym hyponym hyponym | 75 |
| similar_to similar_to | 59 |
| derivat derivat | 59 |
| part_meronym | 49 |
| hyponym derivat | 46 |
| hypernym derivat | 42 |
| derivat hyponym | 40 |
| hypernym hyponym derivat | 37 |
| domain_TOPIC domain_member_TOPIC | 36 |
| derivat hypernym hyponym | 35 |
| also_see | 35 |

Table 1: The most frequent relations chains between a seed and the target.

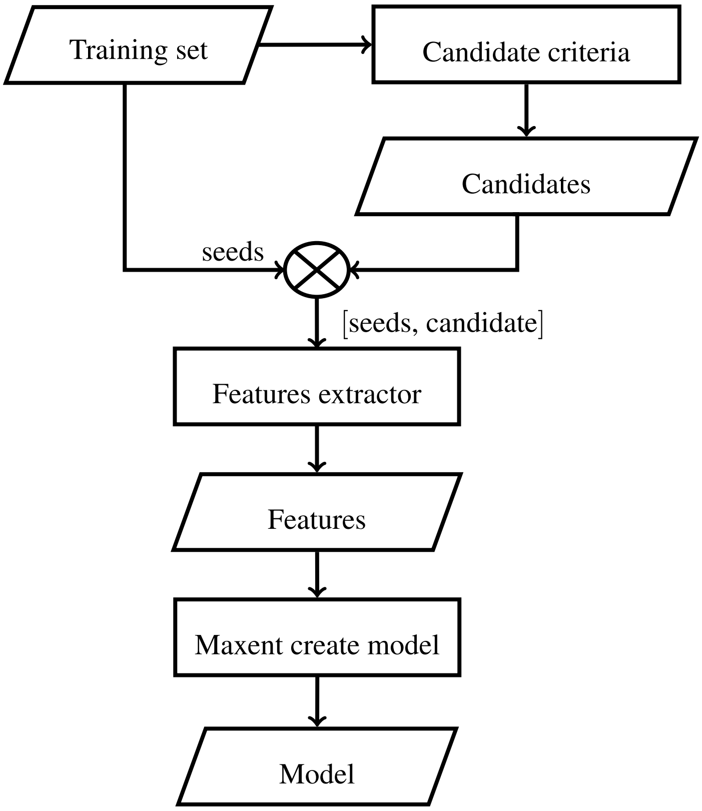

Figure 1: The training flowchart.

Novischi, 2002), who top rank them in finding paths between related concepts for a Question Answering system.) However, they introduce a lot of noise, too, especially when the last relation in the chain is hyponymy and the node from which it starts is one with very many hyponyms.

### 4.4 The system in the shared task competition

We reformulated this as a classification problem. Assuming that having a list of seeds and the list of their possible candidates, the problem will be solved by considering the most probable candidate as the closest to all seeds. We chose `valid` and `invalid` as classification categories.

The system uses the machine learning technique called Maximum Entropy Modeling (MaxEnt for short) and the features needed by MaxEnt are extracted from the kinds of relations presented above, in subsection 4.2. In other words, we mapped each kind of relation to a feature. The entire process has two distinct phases: training and prediction.

The training mechanism is presented in Figure 1. For each training set entry (i.e. the list of 5 seeds and the expected target) a list of possible candidates is generated using the PWN relations chains presented above. We called this process Candidate Criteria. Combining each set of seeds with their candidates we extracted the list of features needed to enter into the MaxEnt process to create the model. For instance, giving the sequence of seeds `away fonder illness leave presence` and two possible candidates `absence` and `being` we obtained the following lists of features ending with the corresponding classification category:

```
domain=s_factotum domain=t_factotum src=1 wn=an wn=he_he_ho_ho
wnshort=he_ho valid

domain=s_factotum domain=t_factotum src=1 wn=he_ho_d_d invalid
```

The following list of features were used:

- wn=*chain*: *chain* represents the relations chain found between any seed and the current candidate. We used short forms to label relations: for example, an stands for antonymy, he for hypernymy, ho for hyponymy, d for derivational relation;

- form=first_upper when at least one seed and the candidate begin with a capital letter; we did not allow for candidates with initial capital letter unless at least one seed had an initial capital letter;

- src=*n* marks the number *n* of seeds that reached the candidate using the PWN chains. In the case of the seed presence and candidate absence there are two chains linking the two words: an and he_he_ho_ho and only presence contributes to them;

- gloss=*n* marks the number *n* of seeds that occur in the target gloss;

- n2gram=high used when any seed occurs in any Google 2-grams with the candidate;

- domain=s_*domain* used to mark the seed domain(s);

- domain=t_*domain* used to mark the candidate domain(s);

- wnshort=*short_chain* here the *short_chain* represents a reduced version of the PWN chain. For example, the chain he_he_ho_ho can be reduced to he_ho (or to a co-hyponym relation, in an extended meaning). The reason is to create an invariant chain that can hold irrespectively of the number of similar consecutive relations. This is useful in hierarchies involving many scientific or artificial nodes which are not known or simply disregarded by common speakers. For example, the chain between hippopotamus and animal is 7 hyponyms long in PWN, whereas for a speaker they are in a direct relation.

The selection of candidates is done using exclusively the PWN relations chains with a maximun length of 5 relations in a chain and only the first literal from the target synset is taken into account (on the assumption that literals PWN synsets are in reverse order of their frequency of occurrence in corpora, with the first as the most frequent). To reduce the number of possible candidates some filtering criteria are applied before pairing them with their corresponding seeds to extract the features described above. These criteria are:

- the candidates that appear among seeds are eliminated;

- the compound terms (recognized by the use of underscore among elements) are excluded;

- the candidates should appear together with any seed among Google 5-grams with a minimum frequency of 5000 (occurrences).

The prediction phase takes the test set and, using the model created in the training phase, produces for each candidate a percent for each category (valid / invalid). The candidate selection and features extraction are done similarly to the training phase. The prediction phase is presented in Figure 2. The result of this phase is a list of candidates (sorted in reverse order) for each set of 5 seeds in the test set. The list of results presented to the shared task organizers contains, for each set of seeds, the best ranked candidate.

### 4.5 Modifications after the competition

After the end of the competition we tried several mechanisms that could improve our results. They were:

- adding two new features that dealt with the part of speech of the words:
  - pos= *s_pos*: the part-of-speech of the seed(s) corresponding to PWN chain that relates to the candidate;
  - pos= *t_pos*: similar for candidate/target;

- considering more literals from synsets when creating the list of candidates.

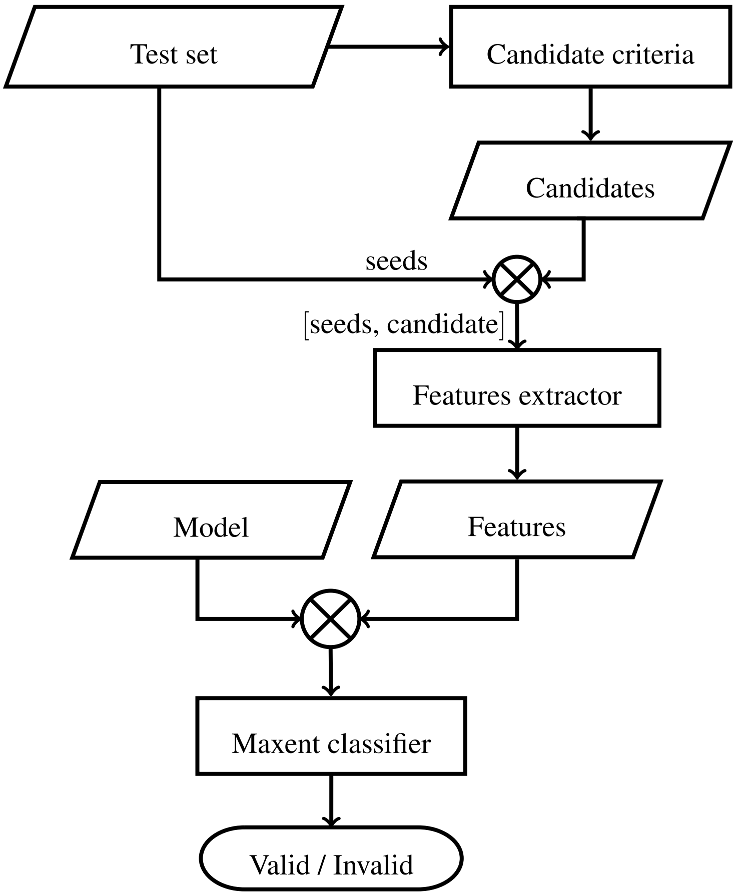

Figure 2: The prediction flowchart.

## 5 Results

### 5.1 Results within the competition

Out of the total number of items (2000) only 30 of our targets matched the ones expected by the organizers, so we obtained 1.50% accuracy.

### 5.2 Improved results after the competition

After considering the part of speech of the words, we were able to match 51 targets, thus increasing the accuracy to 2.55%.

After considering two literals from a synset in the candidates list, the number of matches was 59, so an accuracy of 2.95%.

Furthermore, if we consider the top five candidates in our list, we noticed that 140 targets could be found.

Considering three or even four literals in the synsets did not improve the results (either for the best ranked candidate or for the top 5 ones).

## 6 Conclusions

We presented here the way we dealt with the challenging task proposed by the organizers. Although initially we intended to consider using a large corpus (ukWAC) as well for finding candidates, we found ourselves in the technical impossibility of doing so, because of the costly (timewise especially) resources required by its processing. What is left to be checked is to what extent the lexical and syntactic patterns that can be extracted from a corpus help us improve the results.

We cannot boast good results of our approach mainly because we used only a dictionary (in the form of the PWN). Although it was created on psychological principles about the way words are structured in the speakers' mind, it cannot ensure satisfying results. At least within our approach, the contribution of the relations encoded in PWN is very low. An evaluation of the type n top-ranked candidates could have a higher accuracy for our type of approach. We could dare say that our approach was a further proof of the statement tested by (Zock and Shcwab, 2013): "Words storage does not guarantee their access".

## References

Thorsten Brants, and Alex Franz. 2006. *Web 1T 5-gram Version 1 LDC2006T13*. Philadelphia: Linguistic Data Consortium.

Gemma Bel Enguix, Reinhard Rapp, and Michael Zock. 2014. *How Well Can a Corpus-Derived Co-Occurrence Network Simulate Human Associative Behavior?* Proceedings of the 5th workshop on Cognitive Aspects of Computational Language Learning (CogACLL 2014), pp. 43-48.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Rada Mihalcea, and Dan Moldovan. 2001. *eXtended WordNet: Progress Report*. In Proceedings of NAACL Workshop on WordNet and Other Lexical Resources.

Dan Moldovan, and Adrian Novischi. 2002. *Lexical Chains for Question Answering*. Proceedings of COLING 2002.

Reinhard Rapp. 2008. *The Computation of Associative Responses to Multiword Stimuli*. Proceedings of the workshop on Cognitive Aspects of the Lexicon (CogALex 2008), pp. 102-109.

Michael Zock, and Didier Schwab. 2013. *L'index, une ressource vitale pour guider les auteurs à trouver le mot bloqué sur le bout de la langue*. In Ressources Lexicales : contenu, construction, utilisation, évaluation, N. Gala et M. Zock (eds.). John Benjamins.