A new dynamic approach for lexical networks evaluation

Alain Joubert, Mathieu Lafourcade
LIRMM, Université Montpellier II
{joubert,lafourcade}@lirmm.fr

Abstract

Since September 2007, a large scale lexical network for French is under construction with methods based on popular consensus by means of games (under the JeuxDeMots project). To assess the quality of such a resource built by non-expert users (players of the games), we decided to adopt an approach similar to its construction, that is to say an evaluation by laymen on open class vocabulary. This evaluation is done using a Tip of the Tongue tool.

Keywords: Lexical network, evaluation, JeuxDeMots, TOT software

1 Introduction

Thanks to a significant number of participants in on-line games (JeuxDeMots and PtiClic), we obtained a large scale lexical network for the French language (currently 241000 terms[1] with 1.3 million semantic relations) representing a common general knowledge. So, the community has a lexical resource the quality of which we wish to estimate. A manual evaluation puts forward at least two problems: first, it can be biased by the abilities of the evaluators, and secondly, it may require a prohibitive time as soon as we want to make a consequent evaluation. We might envisage an automatic evaluation against a golden standard, but for French language such a reference does not exist with a similar coverage and relation types. We are facing the difficulties of lexical data evaluation, where no standard reference is available and where a manual evaluation is not possible. More precisely, we would like to answer the following question:

- is our lexical network complete for terms and relations between terms?

Naturally, the realistic answer to this question is negative, in particular because of the evolutionary character of the language. However we can bring out a more practical question:

- for a given term are the relations with the other terms able to characterize it in a unique way?

If the answer is positive, any term may be found via a reduced set of typed clues. A tool helping the resolution of "word on the tip of the tongue" is a way to undertake this evaluation. Through such a tool available on the web, the evaluation can thus be made in a permanent way and with a large number of evaluators (these last ones do not know that they estimate).

We shall begin this paper briefly reminding the TOT problem which introduces our working hypothesis. We shall present then the principles behind our TOT software. In the next section, we shall explicit the realization. Finally, we shall discuss the obtained results to estimate the network quality and we shall see this estimation also allows the acquisition of new relations, enriching thus the (existing) network.

2 The 'tip of the tongue problem' (TOT)

The expression *« it's on the tip of my tongue »* describes a very particular blocking. A speaker trying to express an idea is aware of knowing the term, but he does not manage to produce it, where from the expression, "on the tip of the tongue" (Brown and McNeill, 1966). When we are in a state of TOT we can try to find a term via terms phonologically close (Abrams, 2007), but also via terms having semantic links (Rossi, 2001). To conceive our evaluation tool, we were only interested with semantic associations.

Working hypothesis

The players try to verify the efficiency of the tool: the targeted vocabulary is mainly of low and medium frequency (terms of medium or important difficulty). Given that the vocabulary which activates the TOT and the one which the players of TOT play with are identical, it brings us to postulate that *"the evaluation of a lexical network can be made via a TOT tool or game"*.

We noticed that the motivations of the players consist mainly in trying to trap the tool, either with relatively simple terms and marginal indications, or with rare and recent terms and more direct indications. Thus, we can reasonably conclude that *"the evaluation of a lexical network via a TOT tool results in a pessimistic value of its quality"*. The sampling of the terms to be evaluated is implicitly done by users.

[1] A term can be a compound word (for example: *Eiffel Tower* or *Christmas tree*).

3 Principle of the TOT algorithm

3.1 Principle and realization

The software we developed (called AKI) can be envisaged as a game: the user tries to make the computer 'guess' a target term by supplying, one by one, typed clues[2]. These clues are terms given by the user he thinks they are relevant with the target term. Each of these terms (target and clues) is freely chosen by the user. After each clue, AKI makes the most probable proposition. If it corresponds to the searched target term, the user confirms the proposition as being the proper one; otherwise he introduces a new clue. This dialogue goes on, until either AKI finds the target term, or gives up asking the user to supply the solution. The algorithm relies both on the intersection of sets of terms activated by the clues and the fuzzy set of concepts linked to the clues.

3.2 Algorithm

The precise algorithm is the following. From the first clue i_1, a lexical signature is computed on the basis of what can be found in the lexical network: $S(i_1) = S_1 = t_1, t_2, \ldots$ where the t_i are the terms related to the clue and sorted by descending activation (weight). Put another way, t_1 is the term for which the sum of all relations related to the clue i_1 is the strongest. The first proposition made by AKI, p_1 is this term. The player is supposed to acknowledge it, if it is the target term, otherwise he/she is invited to propose another clue. In this case, the clue and the proposition are removed from the signature: $S'_1 = S_1 - \{p_1, i_1\}$.
With the second clue i_2, the next lexical signature is computed: $S_2 = (S'_1 \cap S(i_2)) - i_2$.The generalized formula at stage n is :

$$S_n = (S'_{n-1} \cap S(i_n)) - i_n \quad \text{and} \quad S'_n = S_n - p_n$$

where i_n is the n-th clue given by the user and p_n la n-th proposition returned by AKI.
With such a process, the size of signatures steadily diminishes with the number of clues. If, the signature becomes empty, then the system has not found the target term. We could stop the process at this stage, but it is more valuable to set a recovering process which will try a simple heuristic. In this case, sum of signatures is made instead of intersections:

$$S_n = (S'_{n-1} + S(i_n)) - i_n \quad \text{and} \quad S'_n = S_n - p_n$$

This recovery leads to a form of learning for the system as if the target term is found this way, unrelated clues are linked in the lexical network. We have found that using the recovering two times before making AKI giving up leads to satisfactory results. Beyond two times, the system tends to propose very general and too loosely related terms (as we can see in the second example showed in figure 1).

4 Realization

4.1 JeuxDeMots[3] : construction of the network

The basic principle leading thanks to an on-line game to the progressive construction of the lexical network, from a pre-existent base of terms, was already described by Lafourcade and Joubert (2010). Let us remind here briefly the progress of a game. A game takes place between two players, in an asynchronous way, based on the concordance of their propositions. When a first player (A) begins a game, an instruction concerning a type of competence (synonyms, opposite, domains …) is displayed, as well as a term T randomly picked in a base of terms. This player A has then a limited time to answer by giving propositions which, to his mind, correspond to the instruction applied to the term T. The number of propositions which he can make is limited inducing players not just type anything as fast as possible, but have to choose amongst all answers he can think of; this limitation increases the relevance of the player's propositions. The same term, along the same instruction, is later proposed to another player B; the process is then identical. For the target term T, we record the common answers from both players. We do not record answers given only by one of the two players but by player pairs. It is a compromise between validating all the answers with necessarily a high noise level and validating by intersection between several players with a reduction of knowledge recovery. The process we perform allows the construction of a lexical network connecting the terms by typed and balanced relations, validated by pairs of players. These relations are labeled by the instruction given to players and they are weighted according to the number of pairs of players who proposed them. Initially, nodes are constituted by an initial set of terms, but if both players in the same game suggest a term initially unknown, it is then added to the lexical base.

Our validation process reminds that one used by von Ahn and Dabbish (2004) for the image indexation or Lieberman and al. (2007) for collection of "common sense knowledge". To our knowledge, it had never been operated in the field of lexical networks construction. In Natural Language Processing, some other Web-based systems exist, such as *Open Mind Word Expert* (Mihalcea and Chklovski, 2003) that aims to create large sense tagged corpora with the help of Web users, or *SemKey* (Marchetti et al., 2007) that exploits WordNet and

[2] We speak about ″typed clues″ because the user can specify a type of relation between each of his clues and his target term (hyperonymy, synonymy, typical localization …) as we will see in section 4.

[3] http://jeuxdemots.org

Wikipedia in order to disambiguate lexical forms to refer to a concept, thus identifying a semantic keyword.

The structure of our network depends on the notions of nodes and relations between nodes, according to a model initially presented by Collins and Quillian (1969), developed in Sowa (1992), and more recently clarified by Polguère (2006). Each node of the network is a lexical item (term or term refinement) connected to other terms via typed and weighted relations corresponding to lexical functions, similar to those of Mel'cuk and al. (1995). When we started JeuxDeMots, in September 2007, the network contained 152 000 terms without any relation. Currently, after approximately 1 000 000 games played by more than 2500 players, our network counts 241 000 terms and more than 1 300 000 relations.

4.2 PtiClic[4] : consolidation of the network

In a similar way to JeuxDeMots (JDM), a second game named PtiClic, presented by Lafourcade and Zampa (2009), takes place between two players in an asynchronous way. A target term T, origin of relations, as well as a cluster of words resulting from terms connected with T in the lexical network produced by JDM are proposed to a first player. Several instructions corresponding to types of relation are also displayed. The player associates words of the cluster with instructions he thinks correspond by a drag- and-dropping. The same term T, as well as the same cluster of words and the same instructions, are also proposed to a second player. According to a principle similar to the one set up for JDM, only the propositions common to both players are taken into account, thus strengthening the relations of the lexical network. Contrary to JDM, the players of PtiClic cannot suggest new terms, but are forced to choose among those proposed. Thus, PtiClic realizes a consolidation of the relations produced by JDM and allows to densify the network.

The collaborative building of resources by non-experts may induce some errors. In fact, as one may expect, we detected some of them, such as classical orthographic mistakes (eg: *théatre* for *théâtre*) or traditional confusions (eg: French singer *Dalida* with the biblical character *Dalila*)… These well-known mistakes are relatively rare and they can be manually detected.

4.3 Term refinement: enrichment of the network

Inspired by the approach developed by Ploux and Victorri (1998) from dictionaries of synonyms, we can determine the various meanings or usages of each term, such as those listed in traditional dictionaries. After validation by an expert lexicographer, we integrate these meanings into the network as refinement nodes of the considered term; the network is so enriched and disambiguated by new nodes and relations. And, of course, the JDM players can create relations from or towards these new nodes.

4.4 AKI[5] : tool of evaluation of the network

The principle of AKI relies on the algorithm presented in section 3. The figure 1 shows two examples of games. Let us note that the users may put before a clue a keyword making reference into semantic functions (hyperonymy, hyponymy, synonymy, antonymy, typical localization …). They correspond to types of relation existing in the JDM network.

Figure 1: Two examples of games.
In the first case, AKI found the target term; player clues and AKI propositions are:
:isa animal → tiger
:loc savannah → feline
neck → giraffe

In the second case, not being able to make any more propositions, AKI gave up and the user furnished the good answer; in this later game, player clues and AKI propositions are:
:isa tent → camping tent
:loc Mongolia → house
round → country
hide → no answer from AKI
The player furnished the answer: *yurt*.

[4] http://pticlic.org

[5] http://www.lirmm.fr/jeuxdemots/AKI.php

5 Evaluation and evolution of the network via AKI

The evaluation, quite as the learning, is made only according to what the players informed. It is thus made on open class vocabulary.

5.1 Quantitative analysis and evolution of the performances

The figure 2 presents the evolution of the ratio between the number of won games and the number of played games on approximately 12000 games currently played.
We analyzed the type of words played by the users. We considered as **common** the words stemming from the Dubois Buyse spelling scale, that is those known by a 12 year-old child. We considered the other ones as **normal**. We noticed that there is a difference of evolution in the improvement of the performances of AKI between the common words and the normal words. Games played with common words didn't show a significant evolution of the obtained results (around 80%). This seems to prove our network is relatively well completed for common words. On the opposite, for normal words, we noticed a slight evolution from less than 60% to approximately 75%. Is it due to an evolution of the players' behaviour? Or is it due to an evolution of our network? We still have to check these hypothesis.

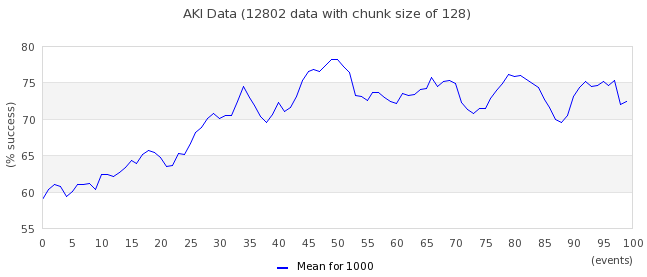

Figure 2: Graph showing the evolution of the ratio between the number of won AKI games and the number of played games (slippery average on the last 1000 played games). The data cover 11835 played games.

Network enrichment: Since the 1st of January 2011, more than 1200 new terms and around 60000 relations have been inserted into the lexical network through AKI. The quasi-totality (90 %) of these terms are named entities (*Jasmine Revolution*) and 10 % of them are compound words and neologisms (*sex by surprise*), often connected to current events.

5.2 Qualitative analysis of the games

Vocabulary type: The vocabulary stands out in number of games played on 24 % of common words, the rest being divided into 50 % of words of medium or low frequency, and 26 % of terms often new and connected to the current events. These later terms often lead to a failure (69 %) which is not surprising because they are new terms or terms with new related clues.

Proposed clues: The average number of clues for finding a word is 2.8. 40 % of common terms are found from the first clue. Less than 3 % of the games are carried on beyond 5 clues.
An analysis of the given clues shows that almost the totality of the games is played "honestly". We can group these clues in two categories:

- **frontal** clues that quickly lead to the solution. In the network, they are strongly connected to the solution.
- **indirect** clues that are weakly connected to the solution and are more strongly connected to other terms.

The games concerning common words correspond to games the typical sequence of which is constituted by a succession of indirect clues; a game with a common target word and frontal clues, such as the first one in figure 1, is only played to discover AKI software: in fact, it is not very funny. The more the target term is rare or recent, the more the typical sequence gets closer to a succession of frontal clues.

5.3 Conclusion of the evaluation

The network allows, for open vocabulary (any term without restriction), to find the term in 75 % of the cases and nearly 80% in the case of common vocabulary. In the case of filtered vocabulary (stemming from the inverted Taboo game[6]), AKI reaches 98,8%, while in this last case an informal evaluation showed that the performance of humans is situated near 80%.

We estimated the performances of five persons on open vocabulary, the given clues being the five terms the most strongly associated in the network. The global performance for people was only 46%.

6 Conclusion

A large scale evolutionary lexical network (JeuxDeMots project) representing a set of common general knowledge is under construction by means of popular consensus. Our purpose is to evaluate the resource quality by means of a TOT tool (named AKI). It allows a broad evaluation of the network in terms of time span, large number of evaluators and wide vocabulary coverage. Whatever the set of the considered terms (common terms or terms of more reduced frequency), the performances of AKI for guessing the proper term are about 75%. The evaluation is

[6] The principle of the Taboo game is to make players guess a target term avoiding five taboo terms. The inverted version of this game consists in guessing the target term furnishing one by one the five taboo terms (which so are clues).

continuous and as the participants are trying to trap AKI, it strengthens the network, but also it increases the strictness of the evaluation - both compensate for each other globally. A question remaining open is to know at which rate of success AKI will asymptotically tend. This value could be an indication of the lexical network quality but also concerning a maximal performance for lexical disambiguation using our resource.

7 References

Abrams L., Trunk D. L. and Margolin S. J. (2007). Resolving tip-of-the-tongue states in young and older adults: The role of phonology. IN L. O. RANDAL (ED.), A*ging and the Elderly: Psychology, Sociology, and Health* (pp. 1-41). HAUPPAUGE, NY: NOVA SCIENCE PUBLISHERS, INC.

von Ahn L., Dabbish L. (2004). Labelling Images with a Computer Game. *ACM Conference on Human Factors in Computing Systems (CHI)*. pp. 319-326

Brown R., McNeill D. (1966). The "tip-of-the-tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5*, pp. 325-337.

Collins A, Quillian M.R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behaviour*, 8(2), pp. 240-248.

Lafourcade M., Joubert A. (2010). Computing trees of named word usages from a crowdsourced lexical network. *Investigationes Linguisticae*, volume XXI, pp. 39-56

Lafourcade M., Zampa V. (2009). « JeuxDeMots and PtiClic: games for vocabulary assessment and lexical acquisition », *Proc. of Computer Games, Multimedia & Allied Technology 09 (CGAT'09)*, Singapore.

Lieberman H., Smith D.A. and Teeters A. (2007). Common Consensus: a web-based game for collecting commonsense goals. *International Conference on Intelligent User Interfaces (IUI'07)*. Hawaï, USA.

Marchetti A., Tesconi M., Ronzano F., Rosella M. and Minutoli S. (2007). SemKey: A Semantic Collaborative Tagging System, *Proceedings of WWW2007*, Banff, Canada.

Mel'čuk I.A., Clas A. and Polguère A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Editions Duculot AUPELF-UREF

Mihalcea R., Chklovski T. (2003). Open Mind Word Expert: Creating Large Annotated Data Collections with Web Users' Help, *Proceedings of the EACL 2003 Workshop on Linguistically Annotated Corpora (LINC 2003)*, Budapest.

Ploux S., Victorri B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement Automatique des Langues*, vol.39/1, 161-182

Polguère A. (2006). « Structural properties of Lexical Systems: Monolingual and Multilingual Perspectives », *Proc. of the Workshop on Multilingual Language Resources and Interoperability* (COLING/ACL 2006), Sydney, pp. 50-59.

Rossi M. (2001). Les lapsus et la production de la parole. *Psychologie Française*, n° 46, pp. 27-41.

Sowa J. (1992). *Semantic networks*, Encyclopedia of Artificial Intelligence, edited by S.C. Shapiro, Wiley, New York.