A Serious Game for Building a Portuguese Lexical-Semantic Network

Mathieu Mangeot
LIG-GETALP & Université de Savoie (France)
41 rue des mathématiques, BP 53
F-38041 GRENOBLE CEDEX 9
FRANCE
mathieu.mangeot@imag.fr

Carlos Ramisch
LIG-GETALP & UFRGS (Brazil)
41 rue des mathématiques, BP 53
F-38041 GRENOBLE CEDEX 9
FRANCE
carlos.ramisch@imag.fr

Abstract

This paper presents a game with a purpose for the construction of a Portuguese lexical-semantic network. The network creation is implicit, as players collaboratively create links between words while they have fun. We describe the principles and implementation of the platform. As this is an ongoing project, we discuss challenges and long-term goals.We present the current network in terms a quantitative and qualitative analysis, comparing it to other resources. Finally, we describe our target applications.

1 Introduction

The creation of lexical resources like wordnets is time consuming and very costly in terms of manpower. Funding agencies and publishing houses are very reluctant to launch new projects. Ironically, in our globalized nowadays world, the need of such resources for communication is growing. In this context, there is hope for building resources via communities of voluntary contributors. But is it possible to use the Wikipedia paradigm for building a rich and broad-coverage lexical resource reusable by humans and machines in NLP projects? Wordnets are very interesting resources, but they suffer of several limitations. First, even if the English wordnet (Miller et al., 1990) is open source and freely available, this is not the case of the EuroWordnets. Second, wordnets and other manually built thesauri are based on linguists' intuition. Information about up-to-date entities (Facebook, Costa Concordia, etc.) and real-world facts are missing. Third, relations between the synsets of wordnets are of limited semantic kinds. We would like to build other relations at the syntactic and lexical level (e.g. collocations).

Our first goal is to build a rich lexical network for the Portuguese language. The relations between nodes (words) is represented in a sophisticated way, by using lexico-semantical functions à la Mel'čuk (Mel'čuk, 1995). The resulting network represents the usage of the language, not the norm. Thus, it may contain frequent spelling mistakes or neologisms. This resource is open-source and freely available. It can be used in several applications: lexicography, printed dictionary, text generation, semantic information extraction, ontology learning, etc. The construction of the resource is done indirectly by contributors through a serious game.

In the next section, the concept of using serious games for building NLP resources will be explained (§ 2). The following section will detail the construction of the Portuguese version of the game (§ 3). Afterwards, we will discuss some preliminary results (§ 4) and finally we present future work (§ 5).

2 Serious Games and NLP

The concept of human contribution, collaboration and computation has been utilized in many applications and scenarios. The work of Luis von Ahn made a breakthrough, especially in ESP game (von Ahn, 2006; von Ahn and Dabbish, 2008). Human computation (crowdsourcing, volunteer contribution) is now seriously considered to be able to solve large computational problems (Speer, 2007). The idea of collecting massive contributions from volunteers through an online game took off recently. Nowa-

days, many serious games or GWAP "Game With A Purpose" (von Ahn and Dabbish, 2008) projects exist in different domains, like Open Mind Common Sense (Singh et al., 2002), ESP games, Learner (Chklovski and Gil, 2005), or CYC project [1]. Concerning more specifically lexical networks, similar projects exist like "small world of words"[2] launched in 2003 by KU Leuwen. For the moment, this project is limited to building relations of only one kind: associated ideas.

Looking at the Wikipedia project, the idea of building lexical resources with the help of communities of voluntary contributors logically comes to mind. Unfortunately, the Wikipedia paradigm cannot be applied to a dictionary. In Wikipedia, articles do not need to follow the same structure, while in a dictionary, the same structure and linguistic theory must be applied to all the articles. Moreover, while it is easy to contribute to an encyclopedia entry, not everyone has the linguistic knowledge to contribute to a dictionary. On reading Wiktionary entries, one quickly realizes that the quality cannot be compared to previous paper dictionaries. Furthermore, there is no such project for bi- and multilingual resources.

When looking at people playing online games through the Internet, one could think that it would be interesting to use this time for playing a game that would build lexical data in the background. In this context, the idea of a serious lexical game emerged. The first version was launched for French in 2007 (Lafourcade and Joubert, 2008). Currently, the French network has approximately 250,000 nodes and 1,330,000 relations.

Our game aims at building a rich and evolving lexical network comparable to the famous English wordnet (Miller et al., 1990). The principle is as follows: a player *A* initiates a game, an instruction is displayed concerning a type of competency corresponding to a lexical relation (e.g. synonym, antonym, domain, intensifier) and a word *W* is chosen randomly in the database. Player *A* has then a limited amount of time for giving propositions that answer the instruction applied to the word *W*.

The same word *W* with the same instruction is proposed to another player *B* and the process is the same. The two half-games of player *A* and player *B* are asynchronous. For each common answer in *A* and *B*'s propositions, the two players earn a certain amount of points and credits. For the word *W*, the common answers of *A* and *B* players are entered into the database. This process participates to the construction of a lexical network linking terms with typed and weighted relations, validated by pairs of players. The relations are typed by the instructions given to the players and weighted with the number of pair players that proposed them. A more detailed description of the game in French is provided by Lafourcade and Zampa (2009).

3 Portuguese Version

The first step was the translation from the French interface by a native Portuguese speaker. Therefore, a preliminary step was to internationalize the text messages allowing for easy translation not only in Portuguese but in any other language. Simultaneously, we developed, and tested an easy step-by-step installer which makes the deployment of the game as easy as installing a CMS software on a server.

A list of seed words must be provided from which the game will chose the proposed terms at the beginning. As the game evolves, people suggest new words not necessarily in the initial dictionary, thus helping the vocabulary to grow. Two resources were used to compose this list of seed words. The first is the DELAS–PB dictionary from NILC (Muniz, 2004). All nouns, verbs, adjectives and adverbs were extracted, resulting in 67,062 words. As these include a large number of rare words, pilot tests showed that the game became annoying when the player ignored the meaning of most of the proposed words. Therefore, the number of Google hits for every word was obtained and only the 20% most common ones were kept, resulting in a list of 13,413 words. To this, the entries of the Brazilian Open Mind Common Sense network (Anacleto et al., 2008) were added. Apertium's lt-toolbox[3] was used in order to obtain the most frequent POS tag for each entry, resulting in 5,129 nouns, 3,672 verbs, 1,176 adjectives, and 201 adverbs. The union with the preceding dictionary resulted in a final seed list of 20,854 words.

[1] http://game.cyc.com/

[2] http://www.smallworldofwords.com/

[3] http://wiki.apertium.org/wiki/Lttoolbox

Once the game is deployed, one of the big challenges is to gather volunteer players. We gave presentations about the game in the academic context and spread the word among Portuguese teachers, arguing that the game could be used to enrich the vocabulary of their students. We also created a Facebook page and linked it in our website. One way to motivate subscribed players to come back is to offer gift words. For the moment, this is done manually but we consider to distribute gifts automatically.

Once the first challenge of gathering a community of players is overcome, the main difficulty is to keep the motivation going. For succeeding, the project needs a person that will animate the community, motivate gamers and publicize the game for recruiting new contributors. Games were launched in other languages, but because of the lack of an animator, they are now in a sleeping state.

Internally, each word is represented as a node in a graph. The directed edges are the lexico-syntactic relations created by the game. Each edge has a type (associated idea, hypernym, hyponym, typical object, etc.) and a weight, corresponding to the number of times the two words co-occurred. Each node has also a weight corresponding to its popularity (proportional to its degree). Part of speech is encoded as edges going from a term to special POS nodes. In addition to the standard attributes, each edge also contains counters that represent the country of players who contributed to its creation. Therefore, we would like to investigate dialectal variations of Portuguese in Portugal, Brasil and other lusophone countries. This information can be important for using the resource in semantic extraction, according to the variation employed in the analyzed text.

4 Preliminary Evaluation

In this preliminary quantitative and qualitative evaluation, we consider only the nodes for which some relation was created, thus excluding all the seed words that were not connected to other words yet.

Figure 1 shows a fragment of the network. Green edges represent associated words, red edges represent hypo- and hypernyms. An inspection of the network shows that most relations created are standard, like *feijó*, *andré*, *amélia* and *jean* are associated with *name*. However, many non-standard relations are also created, like *Cuba* is the antonym of *United States* or *tatoo* is associated to *eternal*. While purists may consider these as noise, we regard it as relations representing not only lexical information but also real-world semantics and language use.

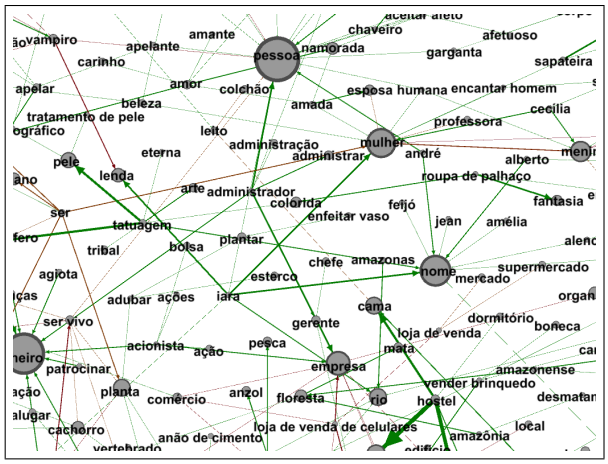

Figure 1: Overview of part of the network.

Currently, the network contains 19,473 word nodes and 20,854 occurrences of POS relations (a word may have several POS). Among those, 347 nodes do not contain POS edges, meaning that they are new additions to the lexicon. A sample of the 20 most popular terms is presented in table 1. They include common hypernym nodes like *thing* and *person* and *animal* but also some everyday language words like *drink*, *car* and *sea*. Only 186 of the current words are multiwords.

From all the nodes in the network, only 1,408 (7.23%) have a degree greater than 1 (excluding POS edges). For the remaining 18,065 nodes, no relation was created for the moment. Since the game was launched three months ago, we expect to obtain more players in order to increase the speed of acquisition. Figure 2 shows user activity in number of games played per day. The curve shows that the

Word	w	Word	w	Word	w
comida	110	hotel	80	pintura	74
***	100	bebida	80	água	72
pessoa	96	mulher	78	porta	72
dinheiro	92	casa	78	mar	72
carne	82	carro	78	empresa	72
nome	80	animal	76	coisa	72

Table 1: Top-20 most connected words and weights (w).

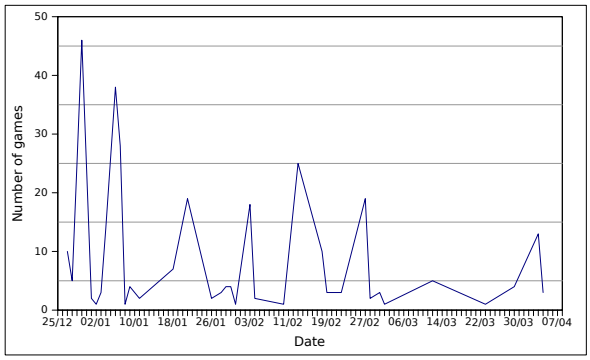

Figure 2: User activity from Dec 27, 2011 to Apr 4, 2012.

number of games is unsteady but it does not seem to increase nor decrease. Analysis of the log files show that players tend to participate a lot and the beginning and then, after one or two weeks, they stop. Thus, it is important not only to attract new players but also to keep them active.

For the moment, the most common edge is of the associated idea type. It corresponds to more than 80% of the edges. Some players bought competences in hypo- and hypernym, which together account for 15.39% of the edges. As these relations are dual, it would be easy to infer new edges.The total number of edges acquired is 1,344.

There is one large connected component in the graph and a large number of small connected components with two or three nodes. The total number of connected components in the graph is 281 (ignoring disconnected nodes), yielding a high modularity of 0.898. The average degree of a node is 0.955, as more than 750 nodes have only one relation and around 200 have 2 relations, and the degree decreases exponentially.

The trend is that, as more relations are added, the current small components will be attached to the larger ones, but also more smaller unconnected graphs will be created. However, we expect that once a large proportion of the nodes has been covered, the network will converge to a single large connected component.

Relation type	**Count**	**%**
Associated	1,126	83.78
Hypernym	115	8.56
Hyponym	81	6.83
Domain	12	0.89
Antonym	10	0.74
Total	1,344	100

Table 2: Number of edges according to types.

For the moment, the coverage of our resource is limited. It only contains as much as 0.91% of the nodes in English Wordnet. As precise numbers about the size of the Portuguese wordnet (Marrafa et al., 2011) are not available, we also queried the online service for the nodes in our network. We found out that 35.87% of the nodes are covered by the Portuguese wordnet. Thus, we believe that collaborative methods can considerably speed up the creation of lexical resources, as in only three months we already have some information complimentary to a 13-years old project.

5 Future work

We presented the deployment and a preliminary evaluation of a game with a purpose that aims at the construction of a Portuguese lexico-semantic network. The coverage of the resource is very limited, but the quality is comparable to existing thesauri and the network keeps constantly growing.

For the moment, we have made available a simple interface in which the user can query for a word and retrieve all the words related to it. For instance, if one searches for the word *loja (store)*, the result is:

- *store* is a *place*
- *cell phone store* is a *store*
- *store* is associated to *buy shirt*
- *store* is associated to *sell toys*
- *store* is associated to *manager*
- *store* is associated to *clothes*

The creation of the network is much less onerous and faster (and more entertaining) than traditional thesauri construction, that can take years of the work of many experts. Our long-term goal is the creation of a large network comparable to existing resources for English. This resource would be extremely useful in many NLP tasks. Once we will have enough data, our goal is to apply it to many other applications like information extraction, word sense disambiguation, semantic inference, textual entailment, etc. This would help to bridge the gap of missing

lexical resources for NLP applications dealing with Portuguese language.

References

Junia Coutinho Anacleto, Aparecido Fabiano P. de Carvalho, Alexandre M. Ferreira, Eliane N. Pereira, and Alessandro J. F. Carlos. 2008. Common sense based applications to advance personalized learning. In *PROC of the IEEE International Conference on Systems, Man and Cybernetics (SMC 2008)*, pages 3244–3249, Singapore.

Timothy Chklovski and Yolanda Gil. 2005. An analysis of knowledge collected from volunteer contributors. In *Twentieth National Conference on Artificial Intelligence (AAAI-05)*, Pittsburgh, Pennsylvania.

Mathieu Lafourcade and Alain Joubert. 2008. Jeuxdemots : un prototype ludique pour l'émergence de relations entre termes. In *JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles*, pages 657–666, Lyon, France, 12-14 mars.

Mathieu Lafourcade and Virginie Zampa. 2009. Jeuxdemots and pticlic: games for vocabulary assessment and lexical acquisition. In *Computer Games, Multimedia & Allied technology 09 (CGAT'09)*, Singapore, 11th-13th May.

Palmira Marrafa, Raquel Amaro, and Sara Mendes. 2011. Wordnet.pt global – extending wordnet.pt to portuguese varieties. In *Proc. of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 70–74, Edinburgh, Scotland, July. ACL.

Igor Mel'čuk. 1995. Lexical functions: A tool for the description of lexical relations in the lexicon. In Leo Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. John Benjamins, Amsterdam/Philadelphia.

G. A. Miller, R. Beckwith, C . Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Marcelo C. M. Muniz. 2004. A construção de recursos linguistico-computacionais para o portugues do brasil: o projeto de unitex-pb. Master's thesis, Instituto de Ciencias Matematicas de Sao Carlos, USP, São Carlos, SP, Brazil.

Push Singh, Thomas Lin, et al. 2002. Open mind common sense: Knowledge acquisition from the general public. In *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*, Irvine, CA, USA.

Robert Speer. 2007. Open mind commons: An inquisitive approach to learning common sense. In *Workshop on Common Sense and Intelligent User Interfaces*, Honolulu, Hawaii, USA., January 28-31.

Luis von Ahn and Laura Dabbish. 2008. General techniques for designing games with a purpose. *Communications of the ACM*, pages 58–67.

Luis von Ahn. 2006. Games with a purpose. *IEEE Computer Magazine*, pages 96–98.