Projet ANR 2015, (Défi 7): LexPro

LEXical resources PROject: strategies, applications and optimal development

Mathieu Lafourcade (Université de Montpellier & LIRMM)

Abstract

Lexical and Semantic Resources are extremely relevant for many Natural Language Processing tasks. Building, improving and linking those resources are a big task on their own and needed for improving results and offering to users easiest and smarter ways to use them. In order to improve the best existent open lexical resources for French, JeuxDeMots, and for Portuguese, OpenWordnet-PT, we propose a project based in three steps: 1. Using strategies of JeuxDeMots for collecting high quality data for OpenWordnet-PT; 2. Linking JeuxDeMots to OpenWordnet-PT, and consequently, to SUMO (Suggested Upper Merged Ontology, the biggest public formal ontology) and Open Multilingual WordNet; 3. Putting fine-grained types into both lexical resources. This final step is an attempt to produce a deeper semantic analysis that is still out of reach using only statistical and probabilistic methods in NLP. To achieve this interdisciplinary project, our proposal gathers linguists, psychologists, logicians and computer scientists.

1 Outline: Lexical Resources, Knowledge Reasoning

This texts presents a 42 months project for studying and share knowledge and skills on Lexical Resources related issues between three institutions: Université de Montpellier / Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), IBM Research - Brazil and the School of Applied Mathematics at Getulio Vargas Foundation (FGV/EMAp).

Since around a decade there has been some growing interest in lexical-semantic networks. It can be explained by the great dependence of many Natural Language Processing (NLP) applications in this kind of lexical resources. The versatility of this kind of structures for knowledge and lexical representation makes them very attractive for a large number of applications (like for instance, machine translation, document indexation, information retrieval, machine learning procedures, knowledge representation and reasoning, etc) the first of them being semantic analysis.

Briefly, a lexical network is a graph in which the nodes are lexical items, and the arcs are the relations between these objects. Resources like Wordnet, Babelnet, ConceptNet or HowNet, which are built on this model are lexical networks. Each lexical resource explores different kind of relations and proposes a particular path to link words and senses. Next we present the network nature of Jeux de Mots, the biggest existent lexical resource for French, and OpenWordnet-PT, the standard version of wordnet for Portuguese.

1.1 The JeuxDeMots Lexical Network

The JeuxDeMots (JDM, [6]) lexical network (developed at LIRMM by the TEXTE team) comprises a hundred binary lexical-semantic relations that can link two terms. These relations are directed and weighted. A relation with a negative weight is considered false (or inhibitory). In October 2015, the network represents 575,781 terms linked with over 23 million lexical and semantic relations (including 218,084 inhibitory relations). This enlarges continuously, fed by many GWAPs and other crowdsourcing activities. Over 11,000 polysemous terms are refined in 33,922 meanings and uses. Besides terms, nodes can hold (1) conceptual information that are considered as language independent, (2) variously contextualized terms of concepts (for example tomato [context] botanic or tomato [contact] food), (3) meta-information under the form of annotations (for example, such relations is true but irrelevant, or uncommon, or strongly cultural and subjective). Generally speaking typed and weighted lexical networks hold both strong precision (like symbolic systems or rule based expert systems) and high recall (like in

vectorial models and statistical approaches). Rules, default rules, and exceptions are implementable with a lexical network.

The graph structure of lexical network is suitable for various kinds of inference approaches usually undertaken with propagations. Propagation is the spread of information from nodes and transmitters within the network. Then we can observe how these information activate some nodes. In this kind of works, a working hypothesis is that some conceptual values associated with terms can verify a form of transitivity. This allows us to automatically infer the information of a term if its neighbors are reasonably provided too. Lexical network is fully compatible with Deep Learning and other general classification algorithms. Finally, multilinguality can be conveniently managed in the context of a lexical network. Each language can thus take advantage of the semantic information hold by other language nodes and relation to enhance its own representations.

1.2 OpenWordnet-PT

The OpenWordnet-PT [4], abbreviated as OpenWN-PT, is a wordnet originally developed as a syntactic projection of the Universal WordNet (UWN) [3]. Its long term goal is to serve as the main lexicon for a system of natural language processing focused on logical reasoning, based on representation of knowledge, using an ontology, such as SUMO [8].

OpenWN-PT has been constantly improved through *linguistically motivated* additions and removals, either manually or by making use of large corpora. This is also the case for the lexicon of nominalizations, called NomLex-PT, that is integrated to the OpenWN-PT [2]. One of the features of both resources is to try to incorporate different kinds of quality data already produced and made available for the Portuguese language, independent of which variant of Portuguese one considers.

The philosophy of OpenWN-PT is to maintain a close connection with Princeton's wordnet since this minimizes the impact of lexicographical decisions on the separation or grouping of senses in a given synset. Such disambiguation decisions are inherently arbitrary [5], thus the multilingual alignment gives us a pragmatic and practical solution. It is practical because Princeton WordNet remains the most used lexical resource in the world. It is also pragmatic, since those decisions will be more useful, if they are similar to what other wordnets say. Of course this does not mean that all decisions will be sorted out for us. As part of our processing is automated and error-prone, we strive to remove the biggest mistakes created by automation, using linguistic skills and tools. In this endeavour we are much helped by the linked data philosophy and implementation, as keeping the alignment between synsets is facilitated by looking at the synsets in several different languages in parallel. For this we make use of the Open Multilingual WordNet's interface [1] through links from our interface.

This lexical enrichment process of OpenWN-PT reported in employs three language strategies: (1) translation; (2) corpus extraction; and (3) dictionaries [9, 11]. The essential fact is that given the constant release of new versions of our openWN-PT, we must ensure the quality of the data that we make available. By quality here we mean not only the data content but its encoding consistency.

OpenWN-PT is nowadays considered the most prominent Wordnet for Portuguese. It is the Portuguese Wordnet chose by the Open Multilingual Wordnet Project [1] and the biggest and more complete freely available Portuguese Wordnet recognized by the Global WordNet Association [2]. Moreover, it has been chosen by Google Translate to be used as their source of lexical information for Portuguese [3].

2 Applications

LRs have many kinds of applications as informational retrieval, sentiments analyses and machine translation. Since the nature of OWN-PT and JDM are different, they have different applications. OWN-PT have structured information on linguistic knowledge of senses, while JDM bets on common sense of words and also structured information related to the use of those words or how they are ontologically connected.

From JDM, for example, an ontology for radiology was built following the GWAP strategies, while through OWN-PT machinery we structured the Brazilian Historical and Bibliographical Dictionary (DHBB), created by historians of Getulio Vargas Foundation (FGV).

[1] `http://compling.hss.ntu.edu.sg/omw/`.
[2] `http://globalwordnet.org/wordnets-in-the-world/`.
[3] `http://translate.google.com/about/intl/en_ALL/license.html`

2.1 Timeliness

One can easily observe that in the decade how the interesting for Lexical Resources have increased. A lot of NLP procedures depend (more or less) of this kind or resource. Here we cite the last two years most important conferences on this topic which we have participated:

- Fourth Workshop On Linked Data In Linguistics: Resources and Applications (LDL 2015) — ACL Workshop, Beijing, 2015.
- Named Entities Workshop (NEWS 2015) — ACL Workshop, Beijing 2015
- 5th World Congress and School on Universal Logic (UniLog 2015) — Istanbul, 2015
- International Conference on Recent Advances in Natural Language Processing (RANLP 2015) — Hissar, 2015
- 22ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015) — Caen, 2015
- Type theory and lexical semantics — ESSLLI Workshop, Barcelona, 2015
- Logic and Engineering of Natural Language Semantics 12 (LENLS 12) — Tokyo, 2015
- 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2014) — Kathmandu, 2014
- Ninth International Conference On Language Resources and Evaluation (LREC'14) — Reykjavik, 2014
- 7th Global WordNet Conference — Tartu, 2014
- 19th International Conference on Application of Natural Language to Information Systems (NLDB 2014) —- Montpellier, 2014
- International Semantic Web Conference — Riva Del Garda, 2014
- 10th International Conference On E-Science (10th IEEE) — Guarujá, 2014
- 11th International Workshop of Logic and Engineering of Natural Language Semantics 11 (LENLS11) — Kanagawa, 2014
- 4ème conférence Extraction et Gestion des Connaissances (EGC 2014) — Rennes, 2014
- Éthique et Traitement Automatique des Langues — Paris, 2014
- 6th Workshop On Formal Ontologies Meet Industry (FOMI) — Rio de Janeiro, 2014
- 11th International Conference on Computational Processing of Portuguese Language (PROPOR) — São Carlos, 2014
- Workshop On Tools and Resources For Automatically Processing Portuguese and Spanish (ToRPorEsp) — PROPOR Workshop São Carlos, 2014

For IBM Research, this project could be seen an a continuation of a partnership that has started one year ago with CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico 'National Counsel of Technological and Scientific Development') on funding cognitive computing systems developed in Brazil.

3 Scientific Objectives and Expected Technological Outcomes

We propose an ANR project focused on the exchange of technology, skills and knowledge on Lexical Resources building and improvement. As LR are the most "linguistically" part of NLP systems, we also would like to bring into the discussion the interfaces between NLP and Formal Semantics. We believe that the actual state of art of NLP procedures does not allow to produce a deep semantic analysis and we propose a solution through Formal Semantics. An initial discussion on how linking Formal Semantics and NLP systems can be found at [10].

Many Formal Semantics theories, as Montagovian Generative Lexicon, deeply presented in [12] and firstly introduced in [7], work with fine-grained semantic types, which provide a high accurate and logically structure semantic information. We believe for achieve the next generation on NLP systems, we must have sophisticated semantic relations.

We rely on Formal Semantics to do a deep semantic analysis of expressions.

Among these parameters, we would like to study, in particular:

- Building and exploitation strategies for lexical resources;
- Different kinds of applications for lexical resources;
- How to get a deeper semantic analysis from Formal Semantics studies,
- How to connected two lexical resources that are based on different principles.

Work plan

1. *Putting common sense information on OpenWordnet-PT* as it is the base of the JeuxDeMots knowledge, and other ontological resources, and it is not present in OWN-PT.
2. *Creating and implement a Game with Purpose for improving OpenWordnet-PT*, as gamefication improves the quality of information we have and makes the task less time-consuming. We also believe that we will have different experiences with a French community and a Brazilian community playing this kind of game.
3. *Linking JeuxDeMots to OpenWordnet-PT*, since the linkage between lexical resources makes easier all cross language tasks, besides guarantees the quality of both resources.
4. *Linking JeuxDeMots to SUMO and OpenMultilingual Wordnet*, since OpenWordnet-PT is already mapped into SUMO and OMW, an easy task is to reproduce this mapping into JeuxDeMots, when it will be already linked to OWN-PT. Mapping JDM into OMW will bring to the French resource not only visibility but provides for its users easiest ways to link cross linguistic data. SUMO is the biggest open formal ontology nowadays and, as it is also based on common sense knowledge, we expect to have an interesting product of it.
5. *Evaluate* the quality and usefulness of this linked JeuxDeMots and probably improve SUMO links as well, as JDM information is not automatically extract, we expect that the quality of its information will be higher than we can find in SUMO.
6. *Implement refined semantic types to both lexical resources*, in order to start to produce a deep semantic analyser, logic based and more than the statistical systems we have today.

4 Coherence of the Proposal and Complementarity of the Participants

The network will gather computer scientists, logicians and linguists. More than that, this project aims to connect the very specific speciality of each researcher: logic, type theory and lexical resources. Lafourcade is the inventor of JeuxDeMots, Retoré proposed the type theory we want to use and Rademaker is the responsible for OWN-PT structuring and Real for OWN-PT linguistic data. Engaging those researchers into a single project certainly improves not only the lexical resources, but also the theoretical discussions behind it.

budget in k€

LEASING Site	*Full name*	*Dept.*	*Equipment*	*Staff*	*Outsourcing*	*Operations*	*Internships*	*total per site*
UM	Université de Montpellier	LIRMM	2	108	0	34	6	150
Tutóia	IBM Research Brazil	Natural Resources	15	48	5	70	18	150
		total budget per sort	20	156	5	104	18	**300**

58 k€is a one year post doc
108k€is a PhD (3 years)
internships are for master student about 2k€for 4 months
operations are for project meetings and workshops and for participation to related conferences
outsourcing is for experiments that need crowdsourcing input to have the necessary amount of data collected in a reasonable time

Scientific leader of the proposal

Mathieu Lafourcade — Maître de Conférences HDR, LIRMM — University of Montpellier 2.

Mathieu Lafourcade (age 46) holds a Ph.D. (U. of Grenoble Joseph Fourier) in Computer Science. He teaches programming, compilation and Natural Language Processing at U. of Montpellier. His research in the TEXTE Team focuses on subjects related to lexical-semantic networks, word sense disambiguation, reasoning, crowdsourcing and GWAPs (games with a purpose). He is the inventor of the JeuxDeMots project and platform.

References

[1] Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[2] Livy Maria Real Coelho, Alexandre Rademaker, Valeria de Paiva, and Gerard de Melo. Embedding nomlex-br nominalizations into openwordnet-pt. In Heili Orav, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 7th Global WordNet Conference*, pages 378–382, Tartu, Estonia, jan 2014.

[3] Gerard de Melo and Gerhard Weikum. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA, 2009. ACM.

[4] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. Openwordnet-pt: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, dec 2012. The COLING 2012 Organizing Committee.

[5] Adam Kilgarriff. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113, 1997.

[6] Mathieu Lafourcade. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th International Symposium on Natural Language Processing*, page 7, Pattaya, Chonburi, Thailand, December 2007.

[7] Bruno Mery, Christian Bassac, and Christian Retoré. A Montagovian generative lexicon. In *Formal Grammar*. CSLI, 2007.

[8] Adam Pease and Christiane Fellbaum. Formal ontology as interlingua: the SUMO and WordNet linking project and global WordNet linking project. In *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 2, pages 25–35. Cambridge University Press, 2010.

[9] Alexandre Rademaker, Valeria de Paiva, Gerard de Melo, Livy Maria Real Coelho, and Maira Gatti. Openwordnet-pt: A project report. In Heili Orav, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 7th Global WordNet Conference*, Tartu, Estonia, jan 2014.

[10] Stergios Chatzikyriakidis; Mathieu Lafourcade; Lionel Ramadier and Manel Zarrouk. Type theories and lexical networks: Using serious games as the basis for multi-sorted typed systems. *TYTLES Proceedings - ESSLLI Workshop*, (3), 2015.

[11] Livy Real, Fabricio Chalub, Valeria de Paiva, Claudia Freitas, and Alexandre Rademaker. Seeing is correcting: curating lexical resources using social interfaces. In *Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference on Natural Language Processing of Asian Federation of Natural Language Processing - Fourth Workshop on Linked Data in Linguistics: Resources and Applications (LDL 2015)*, Beijing, China, jul 2015.

[12] Christian Retoré. The Montagovian Generative Lexicon ΛTy_n: a Type Theoretical Framework for Natural Language Semantics. In Ralph Matthes and Aleksy Schubert, editors, *19th International Conference on Types for Proofs and Programs (TYPES 2013)*, volume 26 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 202–229, Dagstuhl, Germany, 2014. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.