

Building and Exploring Semantic Equivalences Resources

Gracinda Carvalho^{1,2,3}, David Martins de Matos^{2,4}, Vitor Rocio^{1,3}

¹Universidade Aberta, ²L2F/INESC-ID Lisboa, ³CITI - FCT/UNL, ⁴Instituto Superior Técnico/UTL

¹Rua da Escola Politécnica, 147, 1269-001 Lisboa - Portugal,

²Rua Alves Redol 9, 1000-029 Lisboa - Portugal

gracindac@uab.pt,david.matos@inesc-id.pt,vjr@uab.pt

Abstract

Language resources that include semantic equivalences at word level are common, and its usefulness is well established in text processing applications, as in the case of search. Named entities also play an important role for text based applications, but are not usually covered by the previously mentioned resources. The present work describes the WES base, Wikipedia Entity Synonym base, a freely available resource based on the Wikipedia. The WES base was built for the Portuguese Language, with the same format of another freely available thesaurus for the same language, the TeP base, which allows integration of equivalences both at word level and entity level. The resource has been built in a language independent way, so that it can be extended to different languages. The WES base was used in a Question Answering system, enhancing significantly its performance.

Keywords: Information Extraction, Information Retrieval; Question Answering; Semantic Information

1. Introduction

Semantic Equivalences resources group together different words or lexical units that have the same or equivalent meaning. Semantic Equivalences have been thoroughly used in searching, where several words with the same meaning may be used alternatively. The search for any of these words allows the retrieval of related information, that would otherwise be missed. This type of electronic resources, for example organized as thesaurus, are usually constructed from previously compiled publications on paper that were the subject of long years of human effort, to which the digital organization adds a lot more of human hours.

Another aspect of semantic equivalences, that is not covered in the above mentioned resources, are those occurring between named entities. Named entities are a common presence in texts, and its correct identification is a crucial task for many text based applications, as information extraction, information retrieval, question answering, summarization, discourse analysis and opinion mining, just to name a few. The same entity can be identified through different names, hence the utility of similar equivalence resources, this turn for names, not words.

In this paper we present the WES base (Wikipedia Entity Synonyms base), a freely available resource built from the Portuguese version of the Wikipedia, in which alternative names for the same entity are grouped together. Although this is an automatically built resource, like the case of a thesaurus, it reflects the information resulting from a large number of human hours. The fact that the information source, Wikipedia, is constantly being updated and edited ensures that the information is extended and enhanced through time.

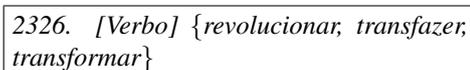
The WES base has a compatible format with another freely available resource for the Portuguese language, the TeP base (Electronic Thesaurus of Portuguese). Together, these two resources can be used to cover synonyms both at word

level and entity name level for Portuguese.

The paper is organised as follows: in Section 2 we briefly describe the TeP base and its structure, and in Section 3 we describe the WES base together with the motivation for its construction its characterisation and results achieved. In Section 4 we describe different approaches used in the construction of named entity resources. We end with Section 5, dedicated to Conclusions.

2. TeP

The TeP base (Dias-da-Silva et al., 2000; Dias-da-Silva et al., 2008; Maziero et al., 2008) is a manually built Thesaurus for the Portuguese Language and is a freely available resource¹. Its structure is based on the WordNet synsets². We used version TeP 2.0 that has around 19 888 synsets, within the morphological classes of *Verbo* [verb], *Substantivo* [noun], *Adjétivo* [adjective] and *Advérbio* [adverb]. An Example of a synset is presented in Figure. 1. In this case it is about verbs with the same meaning as *revolucionar* [revolutionize].



2326. [Verbo] {revolucionar; transfazer; transformar}

Figure 1: Synset of TeP base.

The TeP base includes also information about another semantic relationship, that of antonymy between synsets, or indirect antonymy, indicated at the end of the synset between $\langle \rangle$. An example, in this case for nouns expressing *satisfação* [satisfaction], is given in Figure. 2, with the synset referred as containing nouns with the opposite meaning presented in Figure. 3.

¹<http://www.nilc.icmc.usp.br/tep2/download.htm>

²A synset is an entry representing a semantic equivalence between lexical units or words.

19822. [Substantivo] {agrado, aprazimento, comprazimento, gosto, prazer, satisfação} <18887>

Figure 2: Synset 19822 of TeP base.

18887. [Substantivo] {desagrado, desgosto, desprazer, desprazimento, insatisfação} <19822>

Figure 3: Synset 18887 of TeP base.

3. WES base

The WES base is an entity synonym base that makes use of Wikipedia redirection pages to extract equivalences between entities. The notion of entity in this work is quite wide, and it can be described as a set of words with specific meaning when used together in a fixed order.

3.1. Motivation

Despite its multiple possible usages, the WES base was built to be used in a Question Answering (QA) environment named IdSay (Carvalho et al., 2009; Carvalho et al., 2010).

We proceed to describe briefly this environment, whose architecture is presented in Figure. 4. Each question is treated in the Question Analysis module to determine the question type and other information to be used in the Answer Extraction module (red dashed line in Figure. 4). The Question Analysis module also determines a search string with the information of which words and entities to use in the Document Retrieval module to produce a list of documents that match both. In special cases, for instance definition questions about an entity that has a Wikipedia page, the Set Wikipedia ANswer (SWAN) module produces an answer based on that page, that is directly included in the answers to be treated in the Answer Validation module. In the general case the process proceeds along the blue line in Figure. 4, to the Document Retrieval module, in which the documents of the collection are searched based on the words and entities information derived from the question, producing the list of documents that contain all of them. This list of documents is then processed by the Passage Retrieval module, responsible for the search of passages from the documents that contain the search string, and with length up to a given limit. Passages are extracted in real time, depending on the question information, and may spread across sentence boundaries. The passages are then sent to the Answer Extraction module, where short segments of text (candidate answers) are produced. Finally the answer list is processed to return a ranked list of answers (Answer Validation module). If in one of the steps no data is produced or the results so far are considered unsatisfactory, the search string is revised and the loop starts again, in a process we identify as retrieval cycle (green dashed line in Figure. 4).

Entities play a very important role in QA, for the questions treated by current QA evaluation datasets, such as

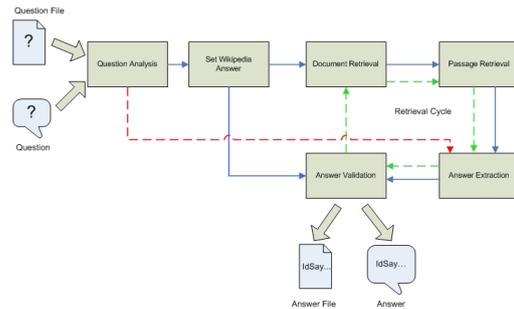


Figure 4: IdSay system architecture.

those of the Text REtrieval Conference (TREC), for the English language, NTCIR Workshop whose main focus is on Asian languages, and the Cross-Language Evaluation Forum (CLEF), QA@CLEF, that since 2004 includes the Portuguese language.

The questions treated by state-of-the-art systems are mostly factoids or definitions that have an entity, the reference entity, that plays a key role in the question, about which a fact is sought for.

The answer is thus a short one, and it can also be considered an entity. Only very few questions in the reference corpus from QA@CLEF for Portuguese have more than one entity present in the question.

Questions addressed by the current state-of-the-art include questions covering factual information of several types (e.g. *Qual é a área da Groenlândia?* [What is the area of Greenland?]), definitions (e.g. *O que é o jagertee?* [What is jagertee?]), list questions (e.g. *Por que estados corre o Havel?* [For which states does the Havel run?]) and cluster questions, i.e. groups of questions linked by anaphoric references (e.g. *Quem foi o criador de Tintin?* [Who was the creator of Tintin?] and *Quando é que ele foi criado?* [When was he created?]), and they may include temporal restrictions (e.g. *Quantos habitantes tinha Berlim em 1850?* [How many inhabitants did Berlin have in 1850?]). The examples given all belong to the Portuguese monolingual task of QA@CLEF 2008.

3.2. Construction

The QA@CLEF data collection for the Portuguese language includes two years of newspaper articles from a Portuguese newspaper and a Brazilian newspaper, together with a frozen version of the Wikipedia, so that results can be reproduced.

The WES base³ was compiled using the Portuguese Wikipedia Version that is part of the text collection of QA@CLEF, that is from November 2006.

Wikipedia combines two characteristics that we consider interesting, one is the fact that it is freely available for public use, and the other is the fact that it results from a collaborative effort from millions of people around the globe, bringing it the benefits of diversity and volume. Both these features contribute for the creation of quality working material.

³Available at the resources section of IdSay web page, <http://www.idsay.net>.

The WES base is obtained from the Wikipedia using content pages names and redirect files. The format is compatible with TeP base, but the equivalences are between entities not lexical units. The synset type, that represents the morphological class of the lexical units in the synset of TeP base, is replaced in the WES base by the label *Wikipedia* which is the source of the information.

The synset information starts by the canonical name of the entity in the Wikipedia (the name of the file that has content) followed by the existing alternative names of the entity (names of redirecting files to the content file). The alternative names are separated by a comma surrounded by spaces, since the names may themselves contain commas, but never surrounded by spaces on both sides.

An Example of a synset is presented in Figure. 5.

15713. [Wikipedia] {*Fernando Henrique Cardoso* , *Presidente Fernando Henrique Cardoso* , *Fernando Henrique* , *FHC*}

Figure 5: Synset of WES base.

The file has 46 586 synsets, ordered alphabetically by canonical names, and the names are kept exactly as they appear in Wikipedia, without any processing.

3.3. Characterisation

The type of equivalences that are represented in the WES base are very encompassing, as is Wikipedia, ranging from abbreviations, acronyms and short names (1), original foreign names (mostly in English, but not only) versus translated names (2), plural and singular (3), scientific names (4), alternative commercial names in different countries (5), alternative spelling variants from European and Brazilian Portuguese (6), capitalization and hyphenization (7), and even spelling mistakes and metonymy (8). Examples of each of these eight types of equivalence are presented in Table 1.

We present the following statistics for the WES base:

- Total number synsets - **46 586**;
- Total number of entities in synsets - **128 724**;
- Mean number of entities per synset - **2.76**;
- Maximum number of entities in a synset - **178**;
- Total number of words entities in synsets - **323 347**;
- Mean number of words in entities - **2.51**;
- Mean number of words in synsets - **6.94**, and
- Maximun number of words in an entity - **22**.

Figure. 6 presents a histogram on the number of synsets per number of entities in synset, for synsets with 10 or less entities. It can be seen that the vast majority of synsets have 2 entities (30 066), and the number of synsets decreases heavily with the number of entities for synset, with the last value in the histogram being 96 for synsets of 10 entities.

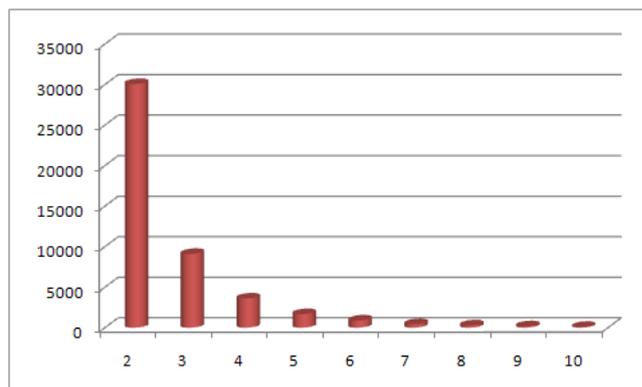


Figure 6: Histogram of synsets by number per entities in synset.

3.4. Usage

The usage of the WES base, together with the TeP base allows the matching to be done using synonyms, not only for words but also at entity level. We describe briefly how the two files are used in IdSay QA system. An entity usually consists of more than a word together that has specific meaning together, respecting word order as is the case of the three words "Fernando Henrique Cardoso", but the alternative name in the synset "FHC" represents the same entity therefore it is added to the entity list at load time, if it is not yet there, being in this case a single word entity. Since entities are made of words, if the word does not yet belong to the word list it is added. At load time of the synset the equivalences structure store information in a single format.

3.5. Results

A baseline of the system was submitted to evaluation at the 2008 edition of QA@CLEF that obtained an accuracy of first answers of 32.5% (Carvalho et al., 2009). In the base line version of the QA system no semantic equivalences were considered. We repeated the tests for a new version of the system with the integration of semantic equivalences (Carvalho et al., 2010). In this later version of the system, Document Retrieval, Passage Extraction, and Answer Extraction modules were adapted to accommodate semantic information, in a way that required no significant loss in efficiency and improving the performance of the system. The results for these two systems on the Portuguese question set of QA@CLEF 2008 are presented in Table 2.

Metric	Baseline	Final
Accuracy at 1st	32.5%	50.5%
Accuracy over 3	42.5%	62.5%
MRR	37.1%	55.5%

Table 2: IdSay Results

The WES base proved to be a valuable resource to the system and helped answering several questions. As an example for the definition question (Question#23 "Quem é FHC?") [Who is FHC?], the equivalence between FHC and "Fernando Henrique Cardoso" of the synset in Figure. 5,

Type	Example WES base entry
1	36224. [Wikipedia] <i>Quod erat demonstrandum</i> , <i>CQD</i> , <i>Q.E.D.</i>
2	30142. [Wikipedia] <i>My Fair Lady</i> , <i>Minha Bela Dama</i> , <i>Minha linda senhora</i>
3	15218. [Wikipedia] <i>Extinção em massa</i> , <i>Extinções em massa</i>
4	25155. [Wikipedia] <i>Lince-ibérico</i> , <i>Cerval</i> , <i>Gato-cerval</i> , <i>Lince da Malcata</i> , <i>Lynx pardinus</i> , ...
5	32290. [Wikipedia] <i>Os Smurfs</i> , <i>Estrunfes</i> , <i>Smurfs</i>
6	39955. [Wikipedia] <i>Sinônimo</i> , <i>Sinónimo</i>
7	42816. [Wikipedia] <i>Tim Berners-Lee</i> , <i>Tim Berners Lee</i>
8	42262. [Wikipedia] <i>The Bionic Woman</i> , <i>A Mulher Biônica</i> , <i>Jamie Sommers</i> , <i>Jaime Sommers</i>

Table 1: Example entries from the WES base corresponding to different types of equivalences.

allowed the retrieval of the correct answer with information on the former Brazilian President.

4. Approaches on Building Named Entities Resources

Exploring existing resources has been a common practice in recent years in the works addressing named entities related tasks. Such are the cases of Named Entities Recognition and Automatic Disambiguation (NER and NED, respectively). The use of Encyclopaedic knowledge, coupled with a variety of different techniques, has been reported in the works of (Bunescu and Paşca, 2006), (Kazama and Torisawa, 2007) for English and (Chrupala and Klakow, 2010) for German.

Other works concentrate on compiling information either from text corpus or encyclopaedic and other sources into a variety of formats that can range from gazetteers, i.e. lists of named entities (Thelen and Riloff, 2002), to more complete ontologies (Suchanek et al., 2007).

Recently, with the increasing interest and usefulness of cross-lingual text treatment, a number of resources have been built that concentrate on the multi-lingual aspects of named entities (Wentland et al., 2008; Steinberger et al., 2011) such as translation and transliteration.

Although these works use the same kind of information and in most cases the same source of knowledge as the present work, Wikipedia, the information is either incorporated in the systems (NER and NED) and not available as autonomous resources, or in different formats from other existing resources and with too less (gazetteers) or too much (full ontologies and multi-lingual resources) information besides the one we propose to address in this work, which is equivalences between entities names.

5. Conclusions

We have presented the WES base, whose objective is to provide a set of alternative names for the same entity, resulting from human intervention, ready to be used in automatic text processing tasks. Its format is simple and straight forward, and allows integration with an existing freely available thesaurus. It was produced for Portuguese, but the uniform structure of the Wikipedia allows it to be produced for other languages, depending only on the source language of the Wikipedia version being processed. It was used in a question answering context, resulting in significant performance improvement.

6. Acknowledgements

The present work was partly supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and FCT project CMU-PT/005/2007.

7. References

- R. Bunescu and M. Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 9–16, Trento, Italy, April.
- G. Carvalho, D. Martins de Matos, and V. Rocio. 2009. IdSay: Question Answering for Portuguese. In *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers, LNCS Series Volume 5706*, pages 345–352. Springer-Verlag, Berlin, Heidelberg. DOI: http://dx.doi.org/10.1007/978-3-642-04447-2_40.
- G. Carvalho, D. Martins de Matos, and V. Rocio. 2010. Improving IdSay: a characterization of strengths and weaknesses in Question Answering systems for Portuguese. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010, LNCS Series Volume 6001*, pages 1–10. Springer-Verlag, Berlin, Heidelberg. DOI: http://dx.doi.org/10.1007/978-3-642-12320-7_1.
- G. Chrupala and D. Klakow. 2010. A named entity labeler for german: Exploiting wikipedia and distributional clusters. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).
- B. Dias-da-Silva, M. F. Oliveira, H. R. Moraes, R. Hasegawa, D. Amorim, C. Paschoalino, and A. C. A. Nascimento. 2000. Construção de um Thesaurus Eletrônico para o Português do Brasil. In *V Encontro para o Processamento computacional da Língua Portuguesa Escrita e Falada, Atibaia, Brasil*, volume 4, pages 1–10.
- B. Dias-da-Silva, A. Di Felippo, and M. G. V. Nunes. 2008. The automatic mapping of princeton wordnet lexical-conceptual relations onto the brazilian portuguese wordnet database. In *Proceedings of the Sixth International*

- Language Resources and Evaluation (LREC'08)*, pages 1535–1541, Marrakech, Morocco, May. European Language Resources Association (ELRA).
- J. Kazama and K. Torisawa. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 698–707, Prague, June. Association for Computational Linguistics.
- E. Maziero, T. Pardo, A. Di Felippo, and B. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 390–392.
- R. Steinberger, B. Pouliquen, M. Kabadjov, J. Belyaeva, and E. van der Goot. 2011. JRC-NAMES: A freely available, highly multilingual named entity resource. In *Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 3230–3237, Hissar, Bulgaria, September.
- F. Suchanek, G. Kasneci, and G. Weikum. 2007. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International World Wide Web conference (WWW 2007)*, pages 697–706, New York, NY, USA. ACM.
- M. Thelen and E. Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 214–221, Philadelphia, USA, June.
- W. Wentland, J. Knopp, C. Silberer, and M. Hartung. 2008. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 3230–3237, Marrakech, Morocco, May. European Language Resources Association (ELRA).