Design and Implementation of a Semantic Search Engine for Portuguese

Carlos Amaral[*], Dominique Laurent[**], André Martins[***], Afonso Mendes[*], Cláudia Pinto[*]

[*]e-mail:{cma, amm, cp}@priberam.pt
[**]e-mail: dlaurent@synapse-fr.com
[***]e-mail: jah@clix.pt

Priberam Informática
Av.ª Defensores de Chaves, 32 – 3º Esq.
1000-119 Lisboa
Tel.: +351 21 781 72 60
Fax: +351 21 781 72 79

Abstract

We present the semantic multilingual question answering engine of the TRUST project, describing its overall architecture, its common multilingual resources, as well as the specific resources, tools and processing mechanisms implemented for the development of the Portuguese language module.

1 Introduction

This paper describes the Portuguese language module developed by Priberam for TRUST[1], *Text Retrieval Using Semantic Technologies*, an EU co-financed project[2], whose aim was the development of a semantic and multilingual search engine capable of processing natural language (NL) questions in English, French, Italian, Polish and Portuguese.

The goal of the search engine is to find a sentence in a set of texts that answers questions in NL. When the user formulates a question, a list of pivots is displayed. For polysemous pivots, a short description of each sense is shown; the user is able to select the sense he finds most adequate, or just accept the one suggested. He may also choose between a local search (i.e., in hard disk) and a Web search.

After the question is submitted, the search engine looks for text blocks containing candidate answers; the selected blocks are ordered by their proximity to the question, and the top ones are passed to the question/answer evaluator. Here, each sentence of each block is parsed and those found the most relevant (based on their likelihood of being real answers) are extracted. Finally, the answers are displayed by descending order of their relevance.

In the next section we explain in detail the various resources developed or used in the project and their role in the Portuguese language module of the search engine. In section 3 we provide an overview of the search engine architecture, namely the *question analysis*, the *indexing process*, the *search procedure* and the *question/answer evaluator*. Finally, in section 4, we present and discuss some experimental results on the performance of the Portuguese search engine, tested both in mono and bilingual environments.

[1] http://www.trustsemantics.com

[2] Cooperative Research (CRAFT) project number IST-1999-56416.

2 Language resources

Under language resources we group together lexical resources (dictionaries, ontologies, etc.) and Natural Language Processing (NLP) tools (morphological and semantic analysers, stemmers, etc.). Additionally, statistical data for morphological and semantic disambiguation, extracted from large corpora[3], turns out to be a very powerful resource, having a crucial influence on the performance of the system.

The current section describes these resources in more detail.

2.1 Ontology

As in other cross-language projects (e.g., Vossen, 1998), an ontology is the common lexical resource of this project. It was initially designed by the French partner, Synapse Développement[4], and then converted into all the languages of the consortium. Our starting point was the translation of the 195 000 English entries of the ontology. We used bilingual (English / Portuguese and French / Portuguese) dictionaries for automatic translation, which covered successfully 25 000 entries. Better results could have been achieved if the dictionaries had a higher number of fixed expressions, which are quite numerous in the ontology. The rest of the translation was done manually.

Briefly, the ontology can be regarded as a lattice whose nodes are the available ontological domains or levels (28 top levels which expand into several sublevels leading to 3387 terminal nodes), and whose entries correspond to the lexical senses of the entries of the lexicon. The organising principle of this ontology consists on building conceptual fields or domains, combining in the same category things that might appear together, such as the actor, the action,

[3] From two major Portuguese newspapers, *Público* (year of 1997) and *Expresso* (from 1997 to 2001).

[4] In fact, the work done by Synapse Développement for *Chercheur,* the French search engine, was the basis for the project.

the instrument, the environment, the place, the manner of acting and so on. The principle also takes into account semantic relations as hyperonymy/hyponymy (e.g., since the noun "climbing" is a kind of sport, it is included in the domain *sports*), as well as proximity relations ("climbing" is also included in the domain *vertical movement*).

2.2 Translation resources

At the end of the ontology translation, each entry in the ontology is a structure {Portuguese word, part-of-speech, sense index, ontological domain, English word} that enables Portuguese-English and English-Portuguese translations through the ontology domain. The following are examples of the entries for the Portuguese verb "comer*"* in the ontology level *everyday life / housing / food / meal*:

{comer, V, 2, 22.2.1.2, eat}
{comer, V, 2, 22.2.1.2, feed}

The combination of the ontologies of all TRUST languages provides a bidirectional word/expression translation mechanism, having English language as an intermediate. This makes possible, for instance, to obtain answers in Portuguese for questions formulated in French, or vice-versa.

2.3 Question categories

Classifying questions into categories is a key task during question analysis (see section 3.1). To address this, a set of 47 question categories has been defined to be shared by all languages of the project.

Question category	Question pattern	Answer pattern	Question category identifer
function	quem + [ser] + function	<FUNC1> + proper noun or named entity	<FUNC1> = profissão, função, etc.
birth date	quando + <BIRTH2> + proper noun or named entity	date + <BIRTH1> or <BIRTH2>	<BIRTH1> = nascimento, etc. <BIRTH2> = nascer, nascido, etc.

Table 1: Examples of question categories.

These categories include, for instance, *denomination* ("Qual era o nome do presidente egípcio quando se deram os ataques terroristas em Luxor?"[5]), *date of event* ("Em que data Jacques Chirac nomeou Jospin primeiro-ministro?"[6]), *town name* ("De que cidade foi Jacques Chirac presidente da câmara?"[7]), *birth date* ("Quando é que nasceu a Dolly?"[8]), *function* ("Quem é o presidente da Albânia?"[9]), etc. For each of the 47 categories, typical question and answer patterns in Portuguese were extracted (see table 1 for some examples).

Question category is the semantic domain of the question, composed by a set of question/answer patterns. Each pattern is defined as a set of words, expressions and ontology domains. At least one question/answer pattern must be present in a sentence for it to be a valid question/answer in the category. We assign to each word in these patterns a *question category identifier*, which will be used to verify if a sentence matches any pattern.

2.4 Lexicon

The underlying architecture of our lexicon database enables the encoding of the data described above, and namely, for each lexical unit:

- Part-of-speech (POS) and further grammatical features;
- Inflections and derivations for each POS;
- Lexical-semantic relations with other words (e.g., synonymy, antonymy, etc.);
- Senses.

Additionally, for each sense, the following information is also encoded:

- A short definition[10];
- Semantic features (e.g., "human subject" for a verb);
- Terminological domains;
- Ontological domains (see section 2.1);
- Equivalents in English for each pair {sense, ontological domain} (see section 2.2);
- A set of identifiers of questions/answers categories where this lexical unit is typically involved (see section 2.3).

2.5 Extraction of named entities, expressions and collocations

Frequently, the questions (as well as the texts to index) have references to named entities (NE), which can be proper nouns, organizations, places, event dates, etc. Besides NE, some expressions (e.g., nominal, adjectival, verbal and adverbial phrases) are frequent enough to justify their handling as if they were single tokens. Indeed, identifying and tagging NE and expressions can strongly improve the final performance of the search engine.

We have implemented a mechanism based on transformation rules (TR), capable of detecting and tagging a large amount of NE and expressions. Some TR were handmade, while others were automatically generated using statistical algorithms over large corpora. Each rule transforms an input sequence of words, lemmas or POS, into a tagged expression.

The NE tagger tries to find a sequence of two or more proper nouns, recognizing it as a single token and classifies the NE thus created according to some criteria, namely the POS established in our lexicon for each element of the entity (e.g., *Luís Vaz de Camões* will be classified as an anthroponym). It also uses groups of conceptually gathered words that will help in the classification of NE: for instance, a sequence of proper nouns preceded by a common noun such as *rio*[11] will be classified as a toponym (e.g., *rio de São Domingos*). Sometimes, for purposes of semantic disambiguation, it

[5] "What was the name of the Egyptian president at the time of the terrorist attacks at Luxor?"
[6] "On what date did Jacques Chirac nominate Jospin prime minister?"
[7] "Of which city was Chirac mayor?"
[8] "When was Dolly born?"
[9] "Who is the president of Albania?"

[10] The definition was encoded semi-automatically (using a Portuguese thesaurus) and manually.
[11] "River"

also takes the context into account, checking what words precede or follow the NE.
Relevant expressions (either for their degree of lexicalization or for their high frequency) are given ontological information, equivalents in English, and, if that is the case, question categories identifiers. This allows such expressions to be processed as if they were single tokens.
In order to verify NE extraction, improve semantic and morphological disambiguation and build TR for fixed expressions, collocations were extracted from corpora, using words, lemmas, POS and ontology levels.

2.6 Morphological and word sense disambiguation

Morphological disambiguation is done in two stages: first, the TR referred in section 2.5 are applied; then, remaining ambiguities are suppressed with a statistical POS tagger based on a second-order hidden Markov model (HMM). This turns out to be a fast and efficient approach using the Viterbi algorithm (see Thede & Harper (1999) and Manning & Schütze (2000) for further details). The prior contextual and lexical probabilities were estimated by processing large, partially tagged corpora. Lexical probabilities are encoded for each *lemma* (rather than for each word). To achieve this, we calculated, for each lemma, its frequency and the relative frequency of its inflections. Then, those lemmas with similar distributions for their inflections were grouped into a smaller number of classes. Clustering techniques based on competitive learning (Haykin, 1994) were used to choose the number of classes, group the lemmas, and characterize each class.
Word sense disambiguation is still at an early stage. Currently, it is implemented as a set of rules that use the semantic features of the lexicon and the ontology domains. We are also investigating the use of collocations with monosemous synonyms of polysemous words for sense disambiguation. Another line of investigation is the extraction of sense selection from parallel corpora in order to build rules for disambiguation.

3 TRUST Search: Portuguese Language Module

Next we describe in further detail the four major tasks involved in our search engine. They are: (i) the question analysis, (ii) the indexing process, (iii) the search procedure, and (iv) the question/answer evaluator.

3.1. Question analysis

For a machine, the simplest method to interpret a NL question is transforming it into a Boolean query by dropping stop-words like "quem", "qual" or "onde"[12] and other frequent words. Although fast, this can be inefficient, since it throws away information that is of great utility for reducing the scope of the question: e.g., dropping down the word "quando"[13] in the sentence "Quando é que nasceu a Dolly?"[14] has the undesirable effect of retrieving answers about *how* and *where*, instead of retrieving only the answers that contain the date of birth. In order to overcome this inconvenient, TRUST defined question categories. During parsing, question category identifiers are traced. Then, these identifiers are compared with typical patterns for each question category; patterns that match correspond to possible categories for the given question (see section 2.3).
After collecting this information, we proceed to the extraction of pivots. Pivots are the key elements of the question, and they can be words, expressions, NE, phrases, numbers, dates, abbreviations, etc. For each pivot, we collect:

- The word or words that make the pivot itself;
- The lemma;
- POS, grammatical and semantic features, obtained after morphological disambiguation;
- Word senses and the index of the sense chosen by the user;
- Ontological and terminological domains for the chosen sense;
- The head of derivation of the pivot with the chosen sense, its equivalent sense if polysemous, its POS, grammatical and semantic features;
- Synonyms of the pivot with the chosen sense, the related senses, their POS and heads of derivation;
- Equivalent pivots in English, their heads of derivation, synonyms and respective heads of derivation.

The data described in this section feeds an information retrieval selector which will rank text blocks.

3.2. Indexing Process

The indexation of each file starts by splitting it in several text blocks; each text block is then parsed, and for each sentence the following information is collected:

- Relevant ontological and terminological domains found in the sentence;
- Question categories for which the sentence may be an answer; these are extracted based on the answer patterns defined in section 2.3.

For each token in the sentence, we collect:

- Flag for stop-word, NE or proper noun;
- POS, grammatical and semantic features;
- Word senses and the index of the selected sense after semantic disambiguation;
- Ontological and terminological domains for the selected sense;
- The head of derivation of the selected sense.

The key elements used for indexation are:

- Words, each represented by a structure {lemma, head of derivation, POS, sense index} (e.g., {ovelha, ovelha, N, 2});
- The relevant ontological domains of each sentence;
- The question categories for which each sentence may be an answer.

Each key element is stored with a pointer to the text block and the file from where it was extracted.

3.3. Search procedure

As said above, the user is allowed to make either local (hard disk) or Web searches. In the first case, a search is made in the index files using as search keys the pivots heads of derivation, their synonyms, the ontological domains and the question categories. In the second case, queries are built from the pivots, considering also their lemmas, inflections and derivations, as well as the synonyms, and submitted to regular search engines (Google, Altavista, Yahoo, etc.). Note that in the second

[12] "who", "what" or "where".
[13] "when".
[14] "When was Dolly born?".

case it is not possible to take profit of POS, ontological and terminological domains, question categories and senses during the search.
In both cases, the output is a set of text blocks submitted to the question/answer evaluator. Each block is then scored taking into account the number and importance of the search keys found in it.

3.4. Question/answer evaluator

The last step consists in analysing the highest ranked text blocks, parsing each sentence and giving it a final score to express its likelihood to answer the question. First, sentences that are found to be below a minimum degree of relevance are excluded. Then, the remaining sentences are scored and kept in an ordered list. The final score is given taking into account the following aspects:

- The number of pivots matching the sentence;
- The number of pivots having in common the lemma or the head of derivation with some token in the sentence;
- The number of pivot synonyms matching the sentence;
- The existence of common question categories between the question and the sentence;
- The number of ontological and terminological domains characterizing the question which are also present in the sentence;
- The score of the block containing the sentence.

Finally, the best sentences are displayed by descendent order of their scores.

4 Experimental Results

The performance of the Portuguese search engine, both in mono and bilingual environments, was tested using large corpora, following the usual procedure in a widely accepted evaluation methodology (Voorhees & Tice, 2000).

Tests	Number of answers							% of correct answer in top 5
	1st	2nd	3rd	4th	5th	Not found[15]	total	
Monolingual	17	6	7	3	3	14	50	72,0%
Bilingual	3	3	1	1	0	17	25	32,0%

Table 2: Experimental results in newspaper corpus.

As for the monolingual test, a list of 50 questions was built manually from the complete editions of the 1997 daily newspaper *Público*. As for the bilingual test, 25 questions were used to retrieve French answers to questions formulated in Portuguese. All the questions seek factual answers (a particular date, fact, location, name, number, etc.), which can be retrieved without inference. We chose to measure the proportion of correct answers in the top five positions (see table 2).

5 Conclusions and future work

We have described the current status of the TRUST Portuguese language module, as well as the results achieved. Based on these results, we conclude that, at the present stage, the whole system performs fairly well in monolingual environment but the overall success is limited in bilingual environment.
Tests show that the system behaves well as far as searching and ranking of documents is concerned. Even so, improvements can be made with better word sense disambiguation, anaphora resolution, increase of the lexical-semantic relations used, more exhaustive semantic feature classification of the lexicon and refined connection between word senses and ontological domains.
Question/answer evaluator still needs improvements. These can be partially achieved by additional syntactic processing both of the questions and of the top ranked sentences in order to avoid answers where there is no syntactic relation between pivots.
The reasons for the low success rate in bilingual environment arise mainly from bad selection of the available translations. They can be overcome with some of the above improvements. The use of parallel corpora for training can also improve cross-language results.

[15] Not found in the top five positions.

6 References

Haykin, S. (1994). Self-Organizing Systems II: Competitive Learning. In *Neural Networks: A Comprehensive Foundation* (pp. 397-443). New York: Prentice Hall.

Manning, C. & Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. Massachusetts: The MIT Press.

Thede, S. M. & Harper, M. P. (1999). A Second-Order Hidden Markov Model for Part-of-Speech Tagging. In *Proceedings of the 37th Annual Meeting of the ACL* (pp. 175-182). Maryland: College Park. Also available at http://acl.ldc.upenn.edu/P/P99/P99-1023.pdf

Voorhees, E. M. & Tice, D. M. (2000). Building a Question Answering Test Collection. In *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 200-207). New York: ACM Press. Also available at http://trec.nist.gov/data/qa/qa_main/qa.ps

Vossen, P. (1998) (ed.). *EuroWordnet: a Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.