Natat in Cerebro: Intelligent Information Retrieval for "The Guillotine" Language Game *

Pierpaolo Basile
Dept. of Computer Science
University of Bari
via E. Orabona, 4
Bari, Italy
basilepp@di.uniba.it

Marco de Gemmis
Dept. of Computer Science
University of Bari
via E. Orabona, 4
Bari, Italy
degemmis@di.uniba.it

Pasquale Lops
Dept. of Computer Science
University of Bari
via E. Orabona, 4
Bari, Italy
lops@di.uniba.it

Giovanni Semeraro
Dept. of Computer Science
University of Bari
via E. Orabona, 4
Bari, Italy
semeraro@di.uniba.it

ABSTRACT

This paper describes OTTHO (On the Tip of my THOught), a system designed for solving a language game, called Guillotine. The rule of the game is simple: the player observes five words, generally unrelated to each other, and in one minute she has to provide a sixth word, semantically connected to the others. The system performs retrieval from several knowledge sources, such as a dictionary, a set of proverbs, and Wikipedia to realize a knowledge infusion process. The main motivation for designing an artificial player for Guillotine is the challenge of providing the machine with the cultural and linguistic background knowledge which makes it similar to a human being, with the ability of interpreting natural language documents and reasoning on their content. Our feeling is that the approach presented in this work has a great potential for other more practical applications besides solving a language game.

1. BACKGROUND AND MOTIVATION

Words are popular features of many games, and they play a central role in many language games. A *language game* is defined as a game involving natural language in which word meanings play an important role. Language games draw their challenge and excitement from the richness and ambiguity of natural language. In this paper we present a system that tries to play the *Guillotine* game. The Guillotine is a language game played in a show on RAI, the Italian National Broadcasting Service, in which a player is given a set of five words (clues), each linked in some way to a specific word that represents the unique solution of the game. She receives one word at a time, and must choose between two different proposed words: one is correct, the other one is wrong. Each time she chooses the wrong word, the prize money is divided by half (the reason for the name *Guillotine*). The five words are generally unrelated to each other, but each of them is strongly related to the word representing the solution. Once the five clues are given, the player has one minute to provide the solution. An example of the game follows: Given the five words *Capital, Pope, City, Colosseum, YellowAndRed*, the solution is *Rome*, because Rome is *Capital* of Italy, the *Pope* lives in Rome, Rome is a *City*, the *Colosseum* is in Rome and *YellowAndRed* is an alternative name for one of the Rome football teams. Often the solution is not so intuitive and the player needs different knowledge sources to reason and find the correct word.

OTTHO (On the Tip of my THOught) tries to solve the final stage of the *Guillotine* game. We assume that the five words are provided at the same time, neglecting the initial phase of choosing the words, that only concerns the reduction of the initial prize.

*The full version appears in [3]

Appears in the Proceedings of the 1st Italian Information Retrieval Workshop (IIR'10), January 27–28, 2010, Padova, Italy.
http://ims.dei.unipd.it/websites/iir10/index.html

2. OTTHO

Guillotine is a *cultural* and *linguistic* game, and for this reason we need to define an extended knowledge base for representing the *cultural* and *linguistic* background knowledge of the player. Next, we have to realize a reasoning mechanism able to retrieve the most appropriate *pieces of knowledge* necessary to solve the game.

2.1 The Knowledge Sources

After a deep analysis of the correlation between the clues and the solution, we chose to include the following knowledge sources, ranked according to the frequency with which they were helpful in finding the solution of the game:
1) Dictionary: the word representing the solution is contained in the description of a lemma or in some example phrases using that lemma;
2) Encyclopedia: as for the dictionary, the description of

an article contains the solution, but in this case it is necessary to process a more detailed description of information;
3) Proverbs and aphorisms: short pieces of text in which the solution is found very close to the clues.

These sources need to be organized and processed in order to model relationships between words. The modeling process must face the problem of the different characteristics of the several knowledge sources, resulting in a set of different heuristics for building the whole model on which to apply the reasoning mechanism. Since we are interested in finding relationships existing between words, we decided to model each knowledge source using the set of correlations existing between terms occurring in that specific source (a proverb, a definition in a dictionary, etc). Indeed, we used a *term-term matrix* containing terms occurring in the modeled knowledge source in which each cell contains the weight representing the degree of correlation between the term on the row and the one on the column. The computation of weights is different for each type of knowledge source.

For the dictionary, we used the on-line De Mauro Paravia Italian dictionary[1], containing 160,000 lemmas. We obtained a lemma-term matrix containing weights representing the relationship between a lemma and terms used to describe it. Because of the general lemma-definition organization of entries in the dictionary, we can fairly claim that the model is language-independent. Each Web page describing a lemma has been preprocessed in order to extract the most relevant information useful for computing weights in the matrix. The text of each Web page is processed in order to skip the HTML tags, even if the formatting information is preserved in order to give higher weights to terms formatted using bold or italic font. Stopwords are eliminated and abbreviations used in the definition of the lemma are expanded. Weights in the matrix are computed using a classical strategy based on a TF-IDF scheme, and normalized with respect to the length of the definition in which the term occurs and the length of the entire dictionary. A detailed description of the heuristics for modeling the dictionary is reported in [5].

As for the dictionary, a TF-IDF strategy has been used for defining the weights in the term-term matrix modeling the knowledge source of proverbs, a collection of $1,600$ proverbs gathered from the web[2].

The process of modeling Wikipedia is different from the one adopted for proverbs and dictionary, due to the huge amount of information to be processed. We adopted a more scalable approach for processing Wikipedia entries, by using models for representing concepts through vectors in a high dimensional space, such as the *Semantic Vectors* or `WordSpace` models [4]. The core idea behind semantic vectors is that words and concepts are represented by points in a mathematical space, and this representation is learned from text in such a way that concepts with similar or related meanings are near to one another in that space (geometric metaphor of meaning). The basis of semantic vectors model is the theory of meaning called *distributional hypothesis*, according to which the meaning of a word is determined by the rules of its use in the context of ordinary and concrete language behavior. This means that words are semantically similar to the extent that they share *contexts* (surrounding words). If 'beer' and 'wine' frequently occur in the same context, say after 'drink', the hypothesis states that they are semantically related or similar.

[1] `http://old.demauroparavia.it/`
[2] `http://web.tiscali.it/proverbiitaliani` and `http://giavelli.interfree.it/proverbi_ita.html`

2.2 The Reasoning Mechanism

We adopt a *spreading activation model* [1], which has been used in other areas of Computer Science such as Information Retrieval [2] as reasoning mechanism for OTTHO. The pure spreading activation model consists of a network data structure of nodes interconnected by links, that may be labeled and/or weighted and usually have directions. In the network for "The Guillotine" game, nodes represent words, while links denote associations between words obtained from the knowledge sources. Spreading in the network is triggered by clues. The activation of clues causes words with related meanings (as modeled in the knowledge sources) to become active. At the end of the weight propagation process, the most "active" words represent good candidates to be the solution of the game.

3. BEYOND THE GAME

The system could be used for implementing an alternative paradigm for *associative retrieval* on collections of text documents [2], in which an initial indexing phase of documents can *spread* further "hidden" terms for retrieving other related documents. The identification of hidden terms might rely on the integration of specific pieces of knowledge relevant for the domain of interest. This might represent a valuable strategy for several domains, such as search engine advertising, in which customers' search terms (and interests) need to be matched with those of advertisers. Spreading activation can be also effectively combined with document retrieval for semantic desktop search.

4. REFERENCES

[1] A. M. Collins and E. F. Loftus. A Spreading Activation Theory of Semantic Processing. *Psychological Review*, 82(6):407–428, 1975.

[2] F. Crestani. Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence*, 11(6):453–482, 1997.

[3] P. Lops, P. Basile, M. de Gemmis, and G. Semeraro. Language Is the Skin of My Thought": Integrating Wikipedia and AI to Support a Guillotine Player. In *AI*IA 2009*, LNCS 5883, pages 324–333. Springer, 2009.

[4] M. Sahlgren. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University, Department of Linguistics, 2006.

[5] G. Semeraro and M. d. G. P. Lops, P. Basile. On the Tip of my Thought: Playing the Guillotine Game. In *IJCAI 2009*, pages 1543–154. Morgan Kaufmann, 2009.