COST Action IC1207 PARSEME meeting, 10–11 March 2014, Athens, WG2: Parsing Techniques for MWEs

Integrating a lexicon-grammar of verbal idioms in a Portuguese NLP system some lexical and parsing issues [1]

J. Baptista[1,3] , N. Mamede[2,3], I. Markov[1,3]
[1] Universidade do Algarve/FCHS/CECL, Faro (Portugal)
[2] Universidade de Lisboa/IST (Portugal)
[3] INESC-ID Lisboa/L2F – Spoken Language Lab (Portugal)

Abstract

Dealing with idioms in Natural Language Processing systems is difficult, among other reasons, because their architecture must be conceived in such a way that it should not preclude the processing of both free word combinations and these, more constraint, expressions. On the other hand, many idioms do have syntactic structure, and can undergo several types of formal variation, thus making them hard to identify in a strictly string pattern-matching approach. Furthermore, many of these expressions are ambiguous between a literal (non-idiomatic) and figurative, non-compositional (idiomatic) use, depending of many linguistic and extra-linguistic factors. This paper presents the way (European) Portuguese verbal idioms have been integrated in fully STRING, a hybrid, statistical and rule-based, natural language processing system, and identify several of the problems that had (and some that still have) to be addressed, in order to adequately identify and process idioms in texts.

1. This paper focuses on *verbal idioms*, e.g. *perder a cabeça*, lit: 'lose the head' (lose one's head), that is, idiomatic (semantically non-compositional) expressions consisting of a verb and at least one constraint argument slot, for which the overall meaning cannot be calculated from the meaning that the individual elements of the expression would present when used independently, in other contexts (M. Gross 1982, 1996).

Extensive lists of verbal idioms, particularly the most frequent ones, have been systematically collected for Portuguese, both the European (Baptista *et al.* 2004, 2005) and the Brazilian (Vale 2001) varieties, along with their main distributional, syntactic and transformational properties, under the Lexicon-Grammar methodological and theoretical framework (M. Gross 1996). Previous studies have shown that the identification of idioms cannot rely neither on strict pattern-matching techniques (Fernandes e Baptista 2007, 2008), nor the use of association measures suffices to identify many idioms (Baptista *et a*l. 2010), hence much manual development of language resources by linguists is required.

In this paper, we address the main issues raised in the process of integrating the lexicon-grammar of European Portuguese verbal idioms into a fully-fledged natural language processing system, STRING (Mamede *et al.* 2012). In order to do so, we briefly present the system in the next section.

2. STRING (`string.l2f.inesc-id.pt`) is a hybrid statistical and rule-based natural language processing chain for Portuguese, with a modular structure, that performs all the basic NLP tasks in four main steps: (i) preprocessing and lexical analysis, (ii) rule-based and (iii) statistical part-of-speech (POS) disambiguation and (iv) parsing. The parsing step is performed by the Xerox Incremental Parser (Ait-Moktar *et al.* 2002), using a rule-based Portuguese grammar jointly developed by the INESC-ID Lisboa and Xerox. XIP first delimits the elementary phrases (or *chunks*, like `NP`, `PP`, *etc*.), and then it extracts the dependencies between the chunk's heads; *e.g*. `SUBJ`ect, `MOD`ifier, `CDIR` (direct complement), *etc*.

3. Considering that idioms have a syntactic structure, STRING's strategy consists in parsing them first as ordinary sentences and only then to identify the word combinations whose meaning is not to be calculated in a compositional way, based on the results of the previous parsing. The idioms are identified by the dependency `FIXED`, which take as its arguments the verb and the frozen elements of the idiomatic expression (the number of arguments depends

[1] This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013 and the European Community Mobility Programme "Erasmus Mundus Action 2, Stand 1" (EMA2), Triple I - Integration, Interaction and Institutions.

on the type of idiom involved). The figure below illustrates the chunking tree and the relevant dependencies extracted for the sentence *O Pedro perdeu a cabeça* (Peter lost his head), where the dependency FIXED has been highlighted:

```
                TOP
     +-----------+----------+
     |           |          |             MAIN(perdeu)
    NP          VF         NP             DETD(Pedro,O)
 +-------+       +      +-------+         DETD(cabeça,a)
 |       |       |      |       |         SUBJ_PRE(perdeu,Pedro)
ART    NOUN    VERB    ART    NOUN        CDIR_POST(perdeu,cabeça)
 +      +-       +     +-       +         FIXED(perdeu,cabeça)
 |      |        |     |        |         NE_PEOPLE_INDIVIDUAL(Pedro)
 O    Pedro   perdeu   a     cabeça       0>TOP{NP{O Pedro} VF{perdeu} NP{a cabeça}}
```

The identification of the idioms uses the previously calculated dependencies and is carried out by a rule like this:

```
        if (VDOMAIN(?,#2[lemma:perder]) & CDIR[post](#2,#3[surface:cabeça])) FIXED(#2,#3)
```

which captures any form of the lemma of the verb (inclunding compound tenses) and the surface form of the direct object (obligatorily after the verb). Around 2,400 rules were semi-automatically built for 10 formal classes of verbal idioms. A list of examples (one for each idiom) was used to test the rules. The remaining of the paper presents the different issues that had to be dealt with in the integration of these rules in a functional module. Finally, we present an estimation of the precision of the system on a large Portuguese corpus of news texts.

References

Ait-Mokhtar, S.; Chanod, J.; Roux, C. (2002). Robustness beyond shallowness: incremental dependency parsing. in Natural Language Engeneering 8-2/3: pp. 121-144.

Baptista, Jorge; Correia, Anabela; Fernandes, Graça (2004). Frozen Sentences of Portuguese: Formal Descriptions for NLP. Workshop on Multiword Expressions: Integrating Processing, International Conference of the European Chapter of the Association for Computational Linguistics, Barcelona (Spain), July 26, 2004. ACL: Barcelona, pp. 72-79.

Baptista, Jorge; Correia, Anabela; Fernandes, Graça (2005). Léxico Gramática das Frases Fixas do Portugués Europeo, in Cadernos de Fraseoloxía Galega 7, pp. 41-53, Santiago de Compostela, Xunta de Galicia/Centro Ramón Piñero para a Investigación en Humanidades.

Baptista, Jorge; Mamede, Nuno; Gomes, Fernando (2010). Auxiliary verbs and verbal chains in European Portuguese. Proceedings of PROPOR'2010. LNCS/LNAI 6001: pp. 110-119. Berlin: Springer.

Baptista, Jorge; Vale, Oto.; Mamede, Nuno (2010). Identificação de expressões fixas em corpora: até onde podem ir os métodos estatísticos? in: Shepherd, T.; Berber Sardinha, T.; Veirano Pinto, M. 2010 (Org.). Caminhos da Linguística de Corpus, Anais do VIII Encontro de Linguística de Corpus (UERJ, 13-14 novembro 2009. Rio de Janeiro, RJ. Mercado de Letras, pp. 159-172.

Fernandes, Graça e Baptista, Jorge (2007). Frozen sentences on large corpus: an experiment. 26th International Colloquium on Compared Lexicon and Grammar, Bonifacio (Corse du Sud), October 2-6, 2007(on-line publication) http://infolingu.univ-mlv.fr/Colloques/Bonifacio/proceedings/fernandes.pdf [2008/09/12].

Fernandes, Graça e Baptista, Jorge (2008). Frozen sentences with obligatory negation: linguistic challenges for natural language processing. in Mellado-Blanco, Carmen (ed.), *Colocaciones y fraseología en los diccionarios*, Frankfurt: Peter Lang, pp.85-96. (paper presented at the International Conference on Phraseology and Paremiology, Santiago de Compostela, Espanha, International Conference On Phraseology And Paremiology, Universidade de Santiago de Compostela, 19-22 de Setembro de 2006).

Gross, Maurice (1996). Lexicon-Grammar, *in* Brown, Keith & Miller, J. (eds.), *Concise Encyclopedia of Syntactic Theories*. Cambridge: Pergamon, pp. 244-259.

Gross, Maurice (1982). Une classification des phrases «figées» du français. *Revue Québécoise de Linguistique*, 12-2, pp.16.

Mamede, Nuno; Baptista, Jorge; Diniz, Cláudio (2012). STRING – A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. in Caseli, H.; Villavicencio, A.; Teixeira, A.; Perdigão, F. (Eds.) Computational Processing of the Portuguese Language, Proceedings of the 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. http://www.propor2012.org/demos.html.

Vale, Oto (2001). Expressões Cristalizadas do Português do Brasil: uma proposta de tipologia. PhD thesis, Universidade Estadual Paulista, Araraquara (SP).