Unsupervised Learning of Multiword Units from Part-of-Speech Tagged Corpora: Does Quantity Mean Quality?

Gaël Dias[1] and Špela Vintar[2]

[1] University of Beira Interior, Computer Science Department,
PT-6200-z001 Covilhã, Portugal
ddg@di.ubi.pt
http://www.di.ubi.pt/~ddg
[2] Faculty of Arts, University of Ljubljana,
SI-1000 Ljubljana, Slovenia
spela.vintar@guest.arnes.si
http://www2.arnes.si/~svinta

Abstract. This paper describes an original hybrid system that extracts multiword unit candidates from part-of-speech tagged corpora. While classical hybrid systems manually define local part-of-speech patterns that lead to the identification of well-known multiword units (mainly compound nouns), we automatically identify relevant syntactical patterns from the corpus. Word statistics are then combined with the endogenously acquired linguistic information in order to extract the most relevant sequences of words. As a result, (1) human intervention is avoided providing total flexibility of use of the system and (2) different multiword units like phrasal verbs, adverbial locutions and prepositional locutions may be identified. Finally, we propose an exhaustive evaluation of our architecture based on the multi-domain, bilingual Slovene-English IJS-ELAN corpus where surprising results are evidenced. To our knowledge, this challenge has never been attempted before.

1 Introduction

Multiword units (MWUs) include a large range of linguistic phenomena, such as compound nouns (e.g. *interior designer*), phrasal verbs (e.g. *run through*), adverbial locutions (e.g. *on purpose*), compound determinants (e.g. *an amount of*), prepositional locutions (e.g. *in front of*) and institutionalized phrases (e.g. *con carne*). MWUs are frequently used in everyday language, usually to precisely express ideas and concepts that cannot be compressed into a single word. As a consequence, their identification is a crucial issue for applications that require some degree of semantic processing (e.g. machine translation, summarization, information retrieval).

In the last 15 years, there has been a growing awareness in the Natural Language Processing (NLP) community of the problems that MWUs pose and the need for their robust handling [1][2]. For that purpose, syntactical [3], statistical [4] and hybrid semantic-syntactic-statistical methodologies [5] have been proposed[1].

[1] We only mention recent works as we assume that the reader is familiar with the field of MWUs extraction.

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 669 – 679, 2005.

However, in the recent past years, the field of MWU acquisition has known a decreasing interest as no new architecture has been proposed that allows the systems to generalize over all MWU linguistic phenomena. In fact, most systems only deal with noun phrases and verb phrases and are defined and tuned for specific languages. In order to avoid these problems and propose more flexible systems, some investigation has been carried out in the field of machine learning but so far with mixed results [6][7][8][9].

In this paper, we propose an original hybrid system called HELAS[2] that extracts MWU candidates from part-of-speech tagged corpora. Unlike classical hybrid systems that manually pre-define local part-of-speech patterns of interest like Noun+Noun, our solution automatically identifies relevant syntactical patterns from the corpus. Word statistics are then combined with the endogenously acquired linguistic information in order to extract the most relevant sequences of words i.e. MWU candidates. Technically, we conjugate the Mutual Expectation (ME) association measure with the acquisition process called GenLocalMaxs [10] in a five step process. First, the part-of-speech tagged corpus is divided into two sub-corpora: one containing only words and one containing only part-of-speech tags. Each sub-corpus is then segmented into a set of positional n-grams i.e. ordered vectors of textual units. Third, the ME independently evaluates the degree of cohesiveness of each positional n-gram i.e. any positional n-gram of words and any positional n-gram of part-of-speech tags. A combination of both MEs is then used to evaluate the global degree of cohesiveness of any sequence of words associated with its respective part-of-speech tag sequence. This combination of MEs is called the Combined Association Measure (CAM). Finally, the GenLocalMaxs retrieves all the MWU candidates by evidencing local maxima of association measure values thus avoiding the definition of global thresholds.

Compared to existing hybrid systems, the benefits of HELAS are clear. By avoiding human intervention in the definition of syntactical patterns, it provides total flexibility of use. Indeed, the system can be used for any language without any specific tuning. HELAS also allows the identification of various MWUs like phrasal verbs, adverbial locutions, compound determinants, prepositional locutions and institutionalized phrases. Finally, it responds to some extent to the affirmation of [11] that claim that "*existing hybrid systems do not sufficiently tackle the problem of the interdependency between the filtering stage [the definition of syntactical patterns] and the acquisition process [the scoring and the election of relevant sequences of words] as they propose that these two steps should be independent*". To our knowledge, no system has ever tried to disclaim this statement.

The paper is divided into four main sections: (1) we present the text corpus segmentation into positional n-grams; (2) we define the Mutual Expectation and the Combined Association Measure; (3) we propose the GenLocalMaxs algorithm as the acquisition process; finally, in (4), we propose an exhaustive evaluation based on the multi-domain bilingual Slovene-English IJS-ELAN corpus [12].

2 Text Segmentation

Positional n-grams are nothing more than ordered vectors of textual units which principles are introduced in the next subsection.

[2] HELAS stands for *Hybrid Extraction of Lexical ASsociations*.