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-6200-001 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].

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.

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

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

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.

2.1 Positional N-grams

The original idea of the positional n-gram model [10] comes from the lexicographic evidence that most lexical relations associate words separated by at most five other words [13]. As a consequence, lexical relations such as MWUs can be continuous or discontinuous sequences of words in a context of at most eleven words (i.e. *5* words to the left of a pivot word, *5* words to the right of the same pivot word and the pivot word itself). In general terms, a MWU can be defined as a specific continuous or discontinuous sequence of words in a (*2.F+1*)-word size window context (i.e. *F* words to the left of a pivot word, *F* words to the right of the same pivot word and the pivot word itself). This situation is illustrated in Figure 1 for the multiword unit Ngram Statistics that fits in the window context of size *2.3+1=7*.

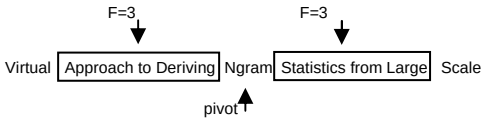

Fig. 1: *7*-word size window context

Thus, any substring (continuous or discontinuous) that fits inside the window context and contains the pivot word is called a positional word n-gram. For instance, the vector [Ngram Statistics] is a positional word n-gram as is the discontinuous sequence [Ngram ___ from] where the gap represented by the underline stands for any word occurring between Ngram and from (in this case, Statistics). Generically, any positional word n-gram may be defined as the following vector of words $[p_{11}\ u_1\ p_{12}\ u_2\ \ldots\ p_{1n}\ u_n]$ where u_i stands for any word in the positional n-gram and p_{1i} represents the distance that separates words u_1 and u_i[3]. Thus, the positional word n-gram [Ngram Statistics] would be rewritten as [0 Ngram +1 Statistics].

[3] By statement, any p_{ii} is equal to zero.

However, in a part-of-speech tagged corpus, each word occurrence is associated to a unique part-of-speech tag. As a consequence, each positional word n-gram is linked to a corresponding positional tag n-gram. A positional tag n-gram is nothing more than an ordered vector of part-of-speech tags exactly in the same way a positional word n-gram is an ordered vector of words. Let's illustrate this situation. Let's consider the following portion of a part-of-speech tagged sentence:

Virtual /JJ Approach /NN to /IN Deriving /VBG Ngram /NN Statistics /NN from /IN Large /JJ Scale /NN Corpus /NN

It is clear that the corresponding positional tag n-gram of the positional word n-gram [0 Ngram +1 Statistics] is the vector [0 /NN +1 /NN]. Generically, any positional tag n-gram may be defined as a vector of part-of-speech tags $[p_{11}\ t_1\ p_{12}\ t_2\ \ldots\ p_{1n}\ t_n]$ where t_i stands for any part-of-speech tag in the positional tag n-gram and p_{1i} represents the distance that separates the part-of-speech tags t_1 and t_i.

So, any sequence of words, in a part-of-speech tagged corpus, is associated to a positional word n-gram and a corresponding positional tag n-gram. In order to introduce the part-of-speech tag factor in any sequence of words of part-of-speech tagged corpus, we present an alternative notation of positional n-grams called positional word-tag n-grams. In order to represent a sequence of words with its associated part-of-speech tags, a positional n-gram may be represented by the following vector of words and part-of-speech tags $[p_{11}\ u_1\ t_1\ p_{12}\ u_2\ t_2\ldots\ p_{1n}\ u_n\ t_n]$ where u_i stands for any word in the positional n-gram, t_i stands for the part-of-speech tag of the word u_i and p_{1i} represents the distance that separates words u_1 and u_i. Thus, the positional n-gram [Ngram Statistics] can be represented by the vector [0 Ngram /NN +1 Statistics /NN] given the text corpus above. This alternative notation will allow us to defining, with elegance, our combined association measure, introduced in the next section.

2.2 Data Preparation

The first step of our architecture deals with segmenting the input text corpus into positional n-grams. First, the part-of-speech tagged corpus is divided into two sub-corpora: one sub-corpus of words and one sub-corpus of part-of-speech tags. The word sub-corpus is then segmented into its set of positional word n-grams exactly in the same way the tagged sub-corpus is segmented into its set of positional tag n-grams.

In parallel, each positional word n-gram is associated to its corresponding positional tag n-gram in order to further evaluate the global degree of cohesiveness of any sequence of words in a part-of-speech tagged corpus. Our basic idea is to evaluate the degree of cohesiveness of each positional n-gram independently (i.e. the positional word n-grams on one side and the positional tag n-grams on the other side) in order to calculate the global degree of cohesiveness of any sequence in the part-of-speech tagged corpus by combining its respective degrees of cohesiveness i.e. the degree of cohesiveness of its sequence of words and the degree of cohesiveness of its sequence of part-of-speech tags. In order to evaluate the degree of cohesiveness of any sequence

of textual units, we use the association measure called Mutual Expectation and introduce the new Combined Association Measure for the specific case of word-tag n-grams.

3. Association Measures

The Mutual Expectation (ME) has been introduced by [10] and evaluates the degree of cohesiveness that links together all the textual units contained in a positional n-gram ($\forall n,\ n \geq 2$) based on the concept of Normalized Expectation and relative frequency. In particular, the ME can be seen as an extension to text data of [14]'s support and confidence measures in the context of association rules: the Normalized Expectation representing the confidence of an association rule and the relative frequency the support of an association rule [10].

3.1 Normalized Expectation

The basic idea of the Normalized Expectation (NE) is to evaluate the cost, in terms of cohesiveness, of the loss of one element in a positional n-gram. In fact, it models an average of a combination of n conditional probabilities present inside a given positional n-gram. Thus, the NE is defined in Equation 1 where the function $k(.)$ returns the frequency of any positional n-gram[4].

$$NE([p_{11}u_1 \ldots p_{1j}u_j \ldots p_{1n}u_n]) = \frac{k([p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n])}{\frac{1}{n}\left(k([p_{22}u_2 \ldots p_{2i}u_i \ldots p_{2n}u_n]) + \sum_{i=2}^{n} k\left(\left[p_{11}w_1 \ldots \hat{p}_{1i}\ \hat{u}_i \ldots p_{1n}u_n\right]\right)\right)} \quad \textbf{(1)}$$

3.2 Mutual Expectation

Many applied works in Natural Language Processing have shown that frequency is one of the most relevant statistics to identify relevant textual associations [15][16]. [10] believes that this phenomenon can be enlarged to part-of-speech tags. From this assumption, he poses that between two positional n-grams with the same NE, the most frequent positional n-gram is more likely to be a relevant sequence. The Mutual Expectation is defined in Equation 2 based on its NE and its relative frequency embodied by the function $p(.)$.

$$ME([p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n]) = p([p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n]) \times NE([p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n]) \quad \textbf{(2)}$$

[4] The "^" corresponds to a convention used in Algebra that consists in writing a "^" on the top of the omitted term of a given succession indexed from 1 to n.

As we said earlier, the ME is going to be used to calculate the degree cohesiveness of any positional word n-gram and any positional tag n-gram. The way we calculate the global degree of cohesiveness of any sequence of words associated to its part-of-speech tag sequence, based on its two MEs, is discussed in the next subsection.

3.3 Combined Association Measure

The drawbacks shown by the statistical methodologies evidence the lack of linguistic information. Indeed, these methodologies can only identify textual associations in the context of their usage. As a consequence, many relevant structures can not be introduced directly into lexical databases as they do not guarantee adequate linguistic structures.

For that purpose, [17] proposed a first attempt to solve this problem without predefining syntactical patterns of interest that bias the extraction process. His idea is simply to combine the strength existing between words in a sequence and the evidenced interdependencies between its part-of-speech tags. We could summarize this idea as follows: the more cohesive the words of a sequence and the more cohesive its part-of-speech tags are, the more likely the sequence may embody a multiword unit.

The degree of cohesiveness of any positional n-gram based on a part-of-speech tagged corpus can then be evaluated by the Combined Association Measure (CAM) defined in Equation 3 where α stands as a parameter that tunes the focus whether on words or on part-of-speech tags.

$$CAM([p_{11}u_1\,t_1 \ldots p_{1i}\,u_i\,t_i \ldots p_{1n}\,u_n\,t_n]) = ME([p_{11}u_1 \ldots p_{1i}\,u_i \ldots p_{1n}\,u_n])^{\alpha} \times ME([p_{11}t_1 \ldots p_{1i}\,t_i \ldots p_{1n}\,t_n])^{1-\alpha} \quad (3)$$

In order to illustrate the *CAM* formula, we illustrate its value for the positional 2-gram [0 Ngram /NN +1 Statistics /NN] in Equation 4.

$$CAM([0\ \text{Ngram}/\text{NN} + 1\,\text{Statistics}/\text{NN}]) = ME([0\ \text{Ngram} + 1\,\text{Statistics}])^{\alpha} \times ME([0/\text{NN} + 1/\text{NN}])^{1-\alpha} \quad (4)$$

We will see in the final section of this paper that different values of α lead to fundamentally different sets of multiword unit candidates. Indeed, α can go from a total focus on part-of-speech tags (i.e. with α=0, the relevance of a word sequence is based only on the relevance of its part-of-speech sequence) to a total focus on words (i.e. with α=1, the relevance of a word sequence is defined only by its word dependencies).

It is important to notice that unlike general smoothing methodologies that use linear interpolation, we preferred, in a first step of our experiments, to use a more drastic smoothing technique. Indeed, with our experience in the field, we believe that radical smoothing could lead to better results than weaker techniques such as linear interpola-

tion. However, we are aware that the linear interpolation should be experimented in further work as a baseline for evaluation. We propose the formula of the linear interpolation in Equation 5.

$$CAM([0\ \text{Ngram}\ /\text{NN}+1\ \text{Statistics}\ /\text{NN}]) = \alpha \times ME([0\ \text{Ngram}+1\ \text{Statistics}]) + (1-\alpha) \times ME([0\ /\text{NN}+1\ /\text{NN}]) \quad (5)$$

Before going to experimentation, we need to introduce the used acquisition process which objective is to extract the MWUs candidates in the overall search space.

4. The Acquisition Process

The GenLocalMaxs [10] proposes a flexible and fine-tuned approach for the selection process as it concentrates on the identification of local maxima of association measure values. So, we may deduce that a positional word-tag n-gram is a MWU if its combined association measure value is higher or equal than the combined association measure values of all its sub-groups of (n-1) words and if it is strictly higher than the combined association measure values of all its super-groups of (n+1) words. Let *CAM* be the combined association measure, *W* a positional word-tag ngram, Ω_{n-1} the set of all the positional word-tag (n-1)-grams contained in *W*, Ω_{n+1} the set of all the positional word-tag (n+1)-grams containing *W* and *sizeof(.)* a function that returns the number of words of a positional word-tag ngram. The GenLocalMaxs is defined as:

$\forall x \in \Omega_{n-1}$, $\forall y \in \Omega_{n+1}$, *W* is a relevant sequence of textual units if

$$(sizeof(W)=2 \ \wedge \ CAM(W) > CAM(y)\ ) \vee (sizeof(W)\neq 2 \ \wedge \ CAM(W) \geq CAM(x) \ \wedge \ CAM(W) > CAM(y))$$

Algo 1. The GenLocalMaxs algorithm

The GenLocalMaxs evidences three interesting properties. First, it allows the testing of various association measures. Second, the GenLocalMaxs allows extracting multiword units obtained by composition. Indeed, as the algorithm retrieves pertinent units by analysing their immediate context, it may identify multiword units that are composed by one or more other MWUs. Third, the GenLocalMaxs shows one important property: it does not depend on global thresholds. A direct implication of this characteristic is the fact that, as no tuning needs to be made in order to acquire the set of all the MWU candidates, the use of the system remains as flexible as possible. Thus, the GenLocalMaxs proposes an excellent evaluation platform for Multiword Unit extraction.

Finally, we propose an exhaustive evaluation of our architecture based on the multi-domain, bilingual Slovene-English IJS-ELAN corpus [12].

5. Evaluation

The main idea of our evaluation is to verify whether our architecture is capable of extending itself to different language families, domains and corpora sizes. For that purpose, we chose three sub-corpora of the multi-domain bilingual Slovene-English IJS-ELAN corpus [12]: the Annex II (Anx2) to the Europe Agreement about EU legislation and politics of 25.000 words, the Slovenian Economic Mirror (Ecmr) about economics of 239.000 words and the Linux Installation and Getting Started (Ligs) about computing of 173.000 words.

In particular, MWUs of sizes 2 to 6 units were extracted from these texts with α ranging from 0.1 to 1[5] and only contiguous units were taken into account.

The evaluation was performed manually by three native speakers of Slovene and two near-native speakers of English, whereby the evaluators were instructed to mark all MWUs belonging to either of the following categories: set phrases, phrasal verbs, adverbial locutions, compound determinants, prepositional locutions and institutionalized phrases, including domain-specific terms and names based on the work developed by [18]. Candidates were marked simply as correct or incorrect with no classes in between. The global precision results are illustrated in Table 1.

Table 1: Average precision

Alpha	English	Slovene
0.1	0.109	0.139
0.2	0.128	0.151
0.3	0.137	0.168
0.4	0.141	0.168
0.5	0.138	0.167
0.6	0.145	0.177
0.7	0.130	0.191
0.8	0.132	0.209
0.9	0.142	0.298
1	0.144	0.284

The overall precision regardless of n-gram type and text type shows that the best result for English is obtained with α = 0.6, while for Slovene the precision seems to be gradually rising as α increases, with the highest value at α = 0.9. The part-of-speech sequence apparently plays a lesser role with a highly inflectional language like Slovene, where on the whole far fewer candidates are extracted due to morphologically reduced frequencies. Although, the results seem to be low, they depend a lot on the corpus size, the type of n-gram and the domain of the corpus. For instance, the best single precision was obtained for Slovene 2grams at α = 0.8 for the smallest corpus, Annex II, and reached 79% precision. It is clear that a deeper analysis needs to be

[5] At the moment of submission, the evaluation for α=0 is still running.

carried out to really understand the behaviour of our system. A complete evaluation over the three corpora is proposed in Table 2.

Table 2: Detailed precision

		English					Slovene				
	Alpha	**2grams**	**3grams**	**4grams**	**5grams**	**6grams**	**2grams**	**3grams**	**4grams**	**5grams**	**6grams**
Anx2	0.1	0.4452	0.2051	0.333	0	0	0.4304	0.1096	0.129	0.0714	0
	0.2	0.4804	0.203	0.5	0	0	0.4336	0.1226	0.1212	0.0909	0.0833
	0.3	0.5166	0.257	0.333	0	0	0.4642	0.1306	0.125	0.166	0.0833
	0.4	0.5298	0.2514	0.333	0	0	0.5168	0.139	0.1142	0.1818	0.0833
	0.5	0.4351	0.2217	0.333	0	0	0.581	0.1616	0.1111	0.1666	0.0833
	0.6	0.5833	0.1947	0.333	0	0	0.6379	0.1666	0.1052	0.1428	0.0833
	0.7	0.4623	0.181	0.25	0	0	0.6595	0.1681	0.1351	0.1428	0.0833
	0.8	0.5	0.188	0.25	0	0	0.7941	0.1637	0.1562	0.1538	0.0833
	0.9	0.5909	0.204	0.25	0	0	0.7037	0.1576	0.1818	0.2	0.07692
	1	0.396	0.1985	0	0	0	0.5714	0.1576	0.2	0.2222	0.07692
Ecmr	0.1	0.1381	0.1308	0	0	0	0.1489	0.1238	0.258	0.125	0
	0.2	0.1432	0.1231	0.0909	0	0	0.1637	0.1337	0.2424	0.125	0
	0.3	0.1597	0.1378	0.1428	0	0	0.1665	0.141	0.2857	0.125	0
	0.4	0.1752	0.1319	0.1463	0	0	0.1801	0.1386	0.2647	0.0625	0
	0.5	0.1764	0.1317	0.1481	0	0	0.2255	0.144	0.2	0	0
	0.6	0.1752	0.1319	0.1463	0	0	0.2866	0.1466	0.2	0	0
	0.7	0.1883	0.132	0.1428	0	0	0.35	0.1555	0.2222	0	0
	0.8	0.1834	0.1336	0.1296	0	0	0.3976	0.1575	0.1904	0	0
	0.9	0.2054	0.1567	0.1228	0	0	0.4015	0.1832	0.2558	0.1538	0.6666
	1	0.2125	0.1865	0.2244	0.1818	0.1	0.4084	0.2012	0.1818	0.1538	0.6666
Ligs	0.1	0.204	0.116	0	0	0.0714	0.159	0.047	0.060	0.108	0.059
	0.2	0.203	0.1603	0.0322	0	0.0714	0.180	0.048	0.058	0.111	0.059
	0.3	0.21	0.2	0.0666	0	0.0714	0.186	0.045	0.061	0.108	0.059
	0.4	0.2084	0.2188	0.0571	0	0.0714	0.183	0.039	0.048	0.111	0.059
	0.5	0.2054	0.2202	0.0957	0.037	0.0666	0.245	0.035	0.054	0.099	0
	0.6	0.2172	0.2086	0.1181	0.0689	0	0.287	0.032	0.049	0.096	0
	0.7	0.2285	0.2032	0.1111	0.0526	0	0.355	0.035	0.047	0.090	0
	0.8	0.2178	0.2046	0.1261	0.0454	0	0.501	0.039	0.045	0.100	0
	0.9	0.1882	0.199	0.1361	0.0847	0	0.462	0.037	0.040	0.099	0
	1	0.2053	0.1933	0.1308	0.0851	0.0487	0.403	0.031	0.033	0.036	0

In comparing overall precision by n-gram type and by text type it becomes clear that the size of the sub-corpus plays a substantial role. The larger the corpus, the lower the precision is, especially for 2-grams. These results are very interesting as it has always been said in the literature that bigger corpora would automatically lead to better results for statistical methodologies. It seems that this assumption does not stand for our

architecture[6]. Indeed, as big corpora evidence large lexical diversity it seems that our system is not as reliable as for small corpora where lexical diversity is small. What could be seen as a problem of scalability is in fact a providential result for many real-world NLP applications which can now integrate a multiword unit recognition "plug-in" that will process texts in real-time[7].

For both English and Slovene, the highest precision is obtained when extracting 2-grams and it then deteriorates with n-gram length, although small differences according to text type may be observed. Moreover, a comparison between English and Slovene shows a constant overall higher precision for Slovene compared to English. The reason for this difference is undoubtedly again the morphological richness of Slovene, which on the one hand results in lower recall, and on the other hand causes for the same phrase to be extracted several times in different cases. However, in order to be extracted at all, an inflected phrase must occur in that form often enough to be spotted, which positively influences precision.

Finally, we evaluated overall precision according to the frequency of the proposed MWUs. As can be expected, precision rapidly increases with frequency, so that for n-grams occurring at least five times, it will almost be increased 50% compared to the precision for n-grams occurring only twice as expressed in Table 3.

Table 3: Overall precision by n-gram frequency

	Slovene					English				
Alpha	**2**	**3**	**4**	**5**	**>5**	**2**	**3**	**4**	**5**	**>5**
0.1	0.175	0.139	0.135	0.217	0.179	0.091	0.098	0.125	0.121	0.134
0.2	0.161	0.152	0.232	0.195	0.196	0.233	0.243	0.292	0.267	0.328
0.3	0.169	0.169	0.243	0.189	0.223	0.257	0.246	0.331	0.268	0.346
0.4	0.159	0.172	0.224	0.191	0.234	0.236	0.203	0.309	0.253	0.320
0.5	0.171	0.203	0.231	0.203	0.270	0.202	0.178	0.328	0.224	0.322
0.6	0.178	0.165	0.195	0.218	0.259	0.198	0.172	0.293	0.232	0.373
0.7	0.173	0.174	0.182	0.200	0.330	0.177	0.177	0.242	0.242	0.300
0.8	0.161	0.197	0.216	0.204	0.323	0.177	0.171	0.247	0.233	0.286
0.9	0.251	0.304	0.335	0.253	0.345	0.192	0.186	0.220	0.215	0.307
1	0.154	0.226	0.232	0.308	0.326	0.192	0.168	0.248	0.216	0.293

6. Conclusion and Future Work

The paper described a system for extracting multiword units from part-of-speech tagged corpora using a hybrid approach that exploits both statistical and linguistic properties. To our knowledge, this experiment had never been attempted before. The

[6] In fact, these results stand for other experiments we did with other corpora and do not only stand for this particular experiment.

[7] In particular, we successfully use this module in our different research works on Topic Segmentation [19] and Web search.

evaluation that was performed for three sub-corpora of a multi-domain Slovene-English corpus shows interesting differences between the languages and between the sub-corpora. The general conclusion is however that the combination of these two layers of information works better than purely statistical methods, while still remaining unsupervised in terms of part-of-speech sequence selection.

Future work will focus on several interesting aspects of language specificity that seem to influence performance. Firstly, the implication that part-of-speech information plays a lesser role for highly inflectional languages like Slovene should be reviewed by expanding the set of languages on the one hand, and by simplifying the tag set on the other hand. We believe that certain layers of the morpho-syntactic analysis, such as gender and number, are redundant for the task at hand.

Secondly, the findings that a smaller corpus yields more accurate MWUs than a larger one, and that frequency nevertheless plays a major role in overall precision, are somewhat controversial and should be explored in more detail. Lexical variation is undoubtedly linked to the corpus composition and corpus homogeneity [19], so that the latter must be considered before any final conclusions can be drawn.

References

1. Tanaka, T. and Baldwin, T.: Noun-Noun Compound Machine Translation: A Feasibility Study on Shallow Processing. In Workshop on Multiword Expressions of the 41st ACL meeting. 7-12 July Sapporo Japan (2003) 17-25.
2. Nivre, J. and Nilsson, J.: Multiword Units in Syntactic Parsing. In: Dias, G., Lopes, J.G.L. and Vintar, S. (eds.): Workshop on Methodologies and Evaluation of Multiword Units in Real-world Applications associated with the 4th International Conference on Languages Resources and Evaluation, Lisbon, Portugal, May 25. ISBN: 2-9517408-1-6. EAN: 0782951740815. (2004) 39-47.
3. Bourigault, D.: Analyse syntaxique locale pour le repérage de termes complexes dans un texte. Traitement Automatique des Langues, vol. 34 (2). (1993) 105-117.
4. Tomokiyo, T. and Hurst, M.: A Language Model Approach to Keyphrase Extraction. In Workshop on Multiword Expressions of the 41st ACL meeting. 7-12 July. Sapporo. Japan. (2003) 33-41.
5. Piao, S., Rayson, P., Archer, D., Wilson, A. and McEnery, T.: Extracting Multiword Expressions with a Semantic Tagger. In Workshop on Multiword Expressions of the 41st ACL meeting. 7-12 July. Sapporo. Japan. (2003) 49-57.
6. Yang, S.: Machine Learning for Collocation Identification. International Conference on Natural Language Processing and Knowledge Engineering, Chengqing Zong (eds), Beijing. China, IEEE Press, October 26-29. ISBN: 0-7803-7902-0. 315-321 (2003)
7. Dias, G. and Nunes, S.: Evaluation of Different Similarity Measures for the Extraction of Multiword Units in a Reinforcement Learning Environment. In M.T. Lino, M.F. Xavier, F. Pereira, R. Costa and R. Silva (eds): Proceedings of the 4th International Conference On Languages Resources and Evaluation, M.T. Lino, M.F. Xavier, F. Pereira, R. Costa and R. Silva (eds), Lisbon, Portugal, May 26-28. ISBN: 2-9517408-1-6. EAN: 0782951740815. (2004) 1717-1721.

8. Díaz-Galiano, M.C, Martín-Valdivia, M.T., Martínez-Santiago, F. and Ureña-López, L.A. Multiword Expressions Recognition with the LVQ Algorithm. In: Dias, G., Lopes, J.G.L. and Vintar, S. (eds.): Workshop on Methodologies and Evaluation of Multiword Units in Real-world Applications associated with the 4th International Conference on Languages Resources and Evaluation, Lisbon, Portugal, May 25. ISBN: 2-9517408-1-6. EAN: 0782951740815. (2004) 12-17.
9. Ogata, T., Terao, K. and Umemura, K.: Japanese Multiword Extraction using SVM and Adaptation. In: Dias, G., Lopes, J.G.L. and Vintar, S. (eds.): Workshop on Methodologies and Evaluation of Multiword Units in Real-world Applications associated with the 4th International Conference on Languages Resources and Evaluation, Lisbon, Portugal, May 25. ISBN: 2-9517408-1-6. EAN: 0782951740815. (2004) 8-12.
10. Dias, G.: Extraction Automatique d'Associations Lexicales à partir de Corpora. PhD Thesis. DI/FCT New University of Lisbon (Portugal) and LIFO University of Orléans (France) (2002).
11. Habert, B. and Jacquemin, C.: Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques. Traitement Automatique des Langues, vol. 34(2). (1993) 5-41.
12. Erjavec, T. The IJS-ELAN Slovene-English Parallel Corpus. International Journal of Corpus Linguistics, 7(1), (2002) 1-20.
13. Sinclair, J.: English Lexical Collocations: A study in computational linguistics. Singapore, reprinted as chapter 2 of Foley, J. A. (ed). 1996, John Sinclair on Lexis and Lexicography, Uni Press. (1974)
14. Agrawal, R., Imielinski, T. and Swami, A.: Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, D.C. USA. (1993) 207--216
15. Justeson, J. and Katz, S.: Technical Terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, vol. 1, (1995) 9-27.
16. Daille, B.: Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. The balancing act combining symbolic and statistical approaches to language, MIT Press, (1996) 49-66.
17. Dias, G.: Multiword Unit Hybrid Extraction. Workshop on Multiword Expressions of the 41st ACL meeting. 7-12 July. Sapporo. Japan. (2003) 41-49.
18. Gross, G.: Les expressions figées en français. Paris, Ophrys. (1996)
19. Dias, G. and Alves, E.: Language-Independent Informative Topic Segmentation. In Proceedings of the 9th International Symposium on Social Communication, Santiago de Cuba, Cuba, January 24-28. (Best Award Paper). ISBN: 959-7174-05-7. (2005). 588-592
20. Kilgarriff, A.: Comparing Corpora. International Jounal of Corpus Lingustics, 6(1), (2001) 97-133.