Corpus Based Validation of WordNet Using Frequency Parameters

Ivan Obradović[1], Cvetana Krstev[2], Gordana Pavlović-Lažetić[3], and Duško Vitas[3]

[1] Faculty of Mining and Geology, Email: ivano@afrodita.rcub.bg.ac.yu
[2] Faculty of Philology, Email: cvetana@matf.bg.ac.yu
[3] Faculty of Mathematics, Email: gordana@matf.bg.ac.yu, vitas@matf.bg.ac.yu
University of Belgrade

Abstract. In this paper we define a set of frequency parameters to be used in synset validation based on corpora. These parameters indicate the coverage of the corpus by wordnet literals, the importance of one sense of a literal in comparison to the others, as well as the importance of one literal in a synset in comparison to other literals in the same synset. The obtained results can be used in synset refinement, as well as in information retrieval tasks.

1 Introduction

The main goal of BalkaNet, the Balkan wordnet project (BWN) is the development of a multilingual database with wordnets for a Bulgarian, Czech, Greek, Romanian, Serbian and Turkish [5]. In its initial phase, Balkanet followed the approach similar to that of EuroWordNet (EWN) developing monolingual wordnets interconnected through an interlingual index (ILI). The development of databases started with a translation of a common set of concepts named Base Concepts in EWN, using the Princeton WordNet (PWN) as the source.

The Serbian WordNet (SWN) has been developed according to this common approach. In the absence of both an explanatory dictionary and an English/Serbian dictionary in electronic form, the translation of English synsets from PWN was done manually, while preserving the PWN semantic structure. The fact that a Serbian dictionary of synonyms does not exist even in paper form made this task even more difficult. In order to establish a relation with the reference six volume explanatory Serbian dictionary of Matica Srpska (RMS), the senses attributed in SWN to literals, or words representing sysnset lemmas in general correspond to the ones in this dictionary. Since the RMS dictionary was published in 1971, new senses had to be attributed in SWN to some of the existing literals but also new literals had to be added. Another reason for refinement of senses defined by RMS is due to the fact that concepts, and hence literal senses in PWN are far more fine grained that the ones in RMS.

The conditions under which SWN has been developed brought up the question of validation of Serbian synsets on corpora. The idea to semantically tag corpora with senses from WordNet has first been realized within the SemCor project [3]. The use of monolingual and multilingual corpora for synset validation leading to the introduction of new literals or removal of existing ones from a synset has already been tackled in [2,4]. Further refinement of this approach is presented in this paper. In order to establish more precise criteria for synset

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 181–186.

validation a set of numerical parameters related to literal-sense pair frequency in corpora has been developed.

2 Frequency Parameters

In order to evaluate the quality of a synset in terms of the comprehensiveness and adequacy of literals used for the lexicalization of a particular concept on one hand, and to establish an ordering among literals within a synset which may be used in information retrieval tasks, on the other, we define a set of indices as numerical measures of relevance of particular literals to synsets they are used in.

Let **S** be the finite set of all synsets within a wordnet: $\mathbf{S} = \{S_i | S_i$ is a synset describing a specific concept, $i = 1, 2, \ldots, N_S\}$; N_S is the total number of synsets within a wordnet. Let **L** be the finite set of all literals used as lexicalizations of one or more concepts: $\mathbf{L} = \{L_k | L_k$ is a literal used in at least one synset of the wordnet, $k = 1, 2, \ldots, N_L\}$; N_L the total number of different literals used within the wordnet. When a literal $L_k \in \mathbf{L}$ is used as a lexicalization of a specific concept described by the synset S_i, it is used in a specific sense (a sense tag is attached to the literal). Omitting the index k of the literal we shall mark all literal-sense pairs within a nonempty synset $S_i \in \mathbf{S}$ in a sequence as $LS_{ij}\,(j = 1, 2, \ldots, n_i)$, where $n_i \geq 1$ is the total number of literals within the nonempty synset S_i.

We shall define the indices for literals within the wordnet, with the aim to determine the relevance of a particular literal to a synset it is used in. In order to determine these indices for a literal a search is performed on a corpus and all occurrences of the selected literal as well as its inflectional forms are identified within a context of a predefined length. We shall first denote the total number of occurrences of a literal L_k within the corpus, regardless of its sense, as L_k^C. The next step is a time-consuming one since it requires manual identification of the sense in which the literal has been used in every particular concordance line identified in the corpus. When this task is completed then the number of occurrences of a literal within the corpus in each specific sense is established. For the senses covered by the wordnet, the appropriate synset S_i the literal belongs to can then be identified. We then proceed taking into account only these senses, and denote the number of times the literal L_k has been used for lexicalization of a concept described by the synset S_i as LS_{ij}^C. The sum of these numbers obtained for all possible senses of a literal covered by the wordnet yields L_k^{WN}, namely, the number of cases when a literal has been used within the corpus as a lexicalization of a concept represented in the wordnet. It is clear that $L_k^{WN} \leq L_k^C$, and that the target of each wordnet should be that for all literals, ideally, $L_k^{WN} = L_k^C$ holds. This would mean that all possible sense usages of a literal identified within the corpus have been covered by wordnet synsets.

If we want to express the relevance of a particular literal L_k to a particular synset S_i within a corpus, then we should compare the number of occurrences of this literal in the corpus denoting the concept represented by the synset S_i, that is LS_{ij}^C, to the total number of occurrences of this literal within the corpus, namely L_k^C. Thus we define the *overall synset relevance index* of a literal as the ratio of the number of times this literal has been used in a specific sense and the total number of occurrences of this literal in the corpus, namely: $I_{ik}^C = LS_{ij}^C / L_k^C$ where the literal from LS_{ij} equals the literal L_k. The index range is $0 < I_{ik}^C \leq 1$, where $I_{ik}^C = 1$ means that the literal L_k is used in one and only one sense, and that is to lexicalize the concept described by the synset S_i.

Since the wordnet coverage of the senses of a literal does not always have to be complete, we define the *wordnet synset relevance index* as the relevance of a particular literal L_k to a particular synset S_i within a more restricted part of the corpus, that is, the part already covered by the wordnet. This index is defined as the ratio of the number of times this literal has been used in a specific sense and the total number of occurrences of a literal within the corpus denoting concepts represented in the wordnet (L_k^{WN}), namely: $I_{ik}^{WN} = LS_{ij}^C / L_k^{WN}$, where the literal from LS_{ij} is the literal L_k. As is the case with I_{ik}^C, the index range is $0 < I_{ik}^{WN} \leq 1$, where $I_{ik}^{WN} = 1$ means that the literal L_k is used in one and only one sense. Since $L_k^{WN} \leq L_k^C$, then $I_{ik}^{WN} \geq I_{ik}^C$. As, ideally, $L_k^{WN} = L_k^C$ should hold for every literal, in an ideal case $I_{ik}^{WN} = I_{ik}^C$ should also be true.

In order to evaluate how close a particular literal L_k is to the ideal case, namely when all its possible senses are covered by the wordnet, we should compare the number of occurrences of a literal within the corpus denoting concepts represented in the wordnet L_k^{WN} to the total number occurrences of the literal within the corpus L_k^C. We therefore define the *wordnet coverage index* of a literal L_k, namely $I_k^{WNC} = L_k^{WN} / L_k^C$. The index ranges between 0 and 1, and in case of full coverage is equal to 1.

All previous indices evaluated the relevance of a literal to a synset regardless of possible other literals within that synset. In order to compare the relevance of a literal within a synset in comparison to other literals denoting the same concept we define the *local synset relevance index* of the literal L_k as the ratio of the number of occurrences of this literal in the corpus denoting the concept represented by the synset S_i, that is LS_{ij}^C, and S_i^C, the number of occurrences of all literals denoting this concept (i.e. belonging to synset S_i): $I_{ik}^L = LS_{ij}^C / S_i^C$, $S_i^C = \Sigma_{j=1}^{n_i} LS_{ij}^C$. It should be noted that the range of the index is $0 < I_{ik}^L \leq 1$ where $I_{ik}^L = 1$, holds when either the synset has only one literal, or other literals from that synset have not appeared in the corpus.

3 The Validation Procedure

In order to test this approach a subset of literal strings, that we called *main strings* has been chosen among those nouns and verbs that have the most senses in Serbian wordnet. Next, a subcorpus has been compiled consisting of contemporary newspaper texts comprising 1.7MW. Concordances were produced for all the inflectional forms of these nouns and verbs. In the next step all the synsets in which the main strings appear have been identified, as well as literal strings, that we called *supporting strings*, that occur beside them in these synsets. For these supporting strings concordances have also been produced. The main and supporting literal strings form the "lexical sample" as defined by the SENSEVAL project [1].

The produced concordances (around 10.000) have than been manually analyzed by lexicographers. In the first step the concordance lines containing the homograph forms have been rejected. In the remaining lines the senses have been identified according to the RMS dictionary and SWN, and marked using the same sense labels.

On the basis of the obtained results tables have been produced and the indices introduced in the section 2 calculated. These data for the noun *lice* and the verb *proizvesti* are given in Tables 1 and 2. For each of the main strings only the senses that are present in SWN are represented. The frequency of occurrence of the these senses in the corpus is given in column

Table 1. The frequency parameters for the lemma *lice* obtained on newspaper corpus

Synset	**lice**	LS_{ij}^C	uloga:1a	lik:3	strana:1b	S_i^C	I_{ik}^C	I_{ik}^{WN}	I_{ik}^L
face, *human face*	1a	33	*	*	*	33	0.063	0.085	1.000
face:6	2a	353	*	*	*	353	0.675	0.912	1.000
character:4, *role:2*, *theatrical role:1*,. . .,	2b	1	34	3	*	38	0.002	0.003	0.026
face:14	3	0	*	*	*	0	0.000	0.000	0.000
side:5,	5a	0	*	*	5	5	0.000	0.000	0.000
	L_k^{WN}	387							
	other	136					0.260	*	*
	L_k^C	523	298	20	861				
	I_k^{WNC}	0.740	I_{ik}^C 0.114	I_{ik}^C 0.150	I_{ik}^C 0.006				
			I_{ik}^L 0.895	I_{ik}^L 0.079	I_{ik}^L 1.000				

LS_{ij}^C. The row L_k^{WN} represents the frequency of the occurrence of all the senses of a string that are covered by SWN, while the row **other** represents the frequency of the occurrence of those senses that are not yet covered. L_k^C is the sum of these two data, and represents the total frequency of the occurrence of the main string, while the index I_k^{WNC} represents their ratio. Among 12 main strings that have been analyzed, three had the value of this index 1, which means that for these strings all the senses identified in RMS dictionary (and perhaps some more) have been included in SWN. For all analyzed literals this index ranges from 0.246 to 1.

Table 2. The frequency indices for the lemma *proizvesti* obtained on newspaper corpus

Synset	**proizvesti**	LS_{ij}^C	prouzrokovati:1	potaknuti:2x	iznedriti:1	proizvoditi:3	napraviti:1a	S_i^C	I_{ik}^C	I_{ik}^{WN}	I_{ik}^L
produce:3,. . .	1a	6	31	1	*	*	*	38	0.090	0.091	0.158
yield:1, *give:2*,. . .	1b	1	*	*	0	*	*	1	0.015	0.015	1.000
produce:2, *make:6*,. . .	3	59	*	*	*	106	21	186	0.881	0.894	0.317
	L_k^{WN}	66									
	other	1							0.015	*	*
	L_k^C	67	31	1	99	114	159				
	I_k^{WNC}	0.985	I_{ik}^C 1.000	I_{ik}^C 1.000	I_{ik}^C 0.000	I_{ik}^L 0.930	I_{ik}^L 0.132				
			I_{ik}^L 0.816	I_{ik}^L 0.026	I_{ik}^L 0.000	I_{ik}^L 0.570	I_{ik}^L 0.113				

The parameter S_i^C gives the overall occurrence of the synset, that is all its literals, in the corpus. The indices I_{ik}^C, I_{ik}^{WN}, and I_{ik}^L in the upper part of the table refer to the main string, while the same indices in the lower part refer to the appropriate supporting strings. The first one is the ratio LS_{ij}^C/L_k^C: for instance, for the sense 1a of the main string *lice*, this index is 0.063, which means that this sense represents 6.3% of all the occurrences of this string in corpus. The second index is the ratio LS_{ij}^C/L_k^{WN}. For the same sense of the string *lice* its value is 0.085 meaning that it covers 8.5% of all the occurrences that represent senses from SWN. Finally, the third index is the ratio LS_{ij}^C/S_i^C. For the sense 2a of the string *lice* the value of this index is 0.026, meaning that of all occurrences of this synset, 2.6% were represented by this particular literal.

If for some string the value of its index I_{ik}^L is close to 0 it can indicate that it has been misplaced in the synset, especially in the cases when both indices L_k^C and S_i^C are considerably greater than 0. For instance, that is the case for the string *napraviti:1a* (Table 2). The string *napraviti* has a considerable frequency on corpus ($L_k^C = 159$), and the synset to which the literal string *napraviti:1a* belongs also has a considerable frequency ($S_i^C = 186$). However, its local synset relevance index is relatively low ($I_{ik}^L = 0.113$), and the synonymy of the literal string *napraviti:1a* with the main string *proizvesti* should be reconsidered.

The calculated indices enable the ordering of the literal strings in a synset. This can be useful for information retrieval (IR) tasks that are seen as one of the most interesting applications of BWN. Especially, strings that have a low value of I^L and a high value of I^C and which are not necessarily misplaced in a synset, can be neglected in IR tasks, thus reducing the recall but improving the precision.

Table 3. The frequency parameters for the lemma *lice* obtained on literary corpus

Synset	**lice**	LS_{ij}^C	uloga:1a	lik:3	strana:1b	S_i^C	I_{ik}^C	I_{ik}^{WN}	I_{ik}^L
face, human face	1a	380	*	*	*	380	0.936	0.979	1.000
face:6	2a	3	*	*	*	3	0.007	0.008	1.000
character:4, role:2,	2b	3	6	1	*	10	0.007	0.008	0.300
face:14	3	0	*	*	*	0	0.000	0.000	0.000
side:5,	5a	2	*	*	4	6	0.005	0.005	0.333
	L_k^{WN}	388							
	other	18					0.044	*	*
	L_k^C	406	22	25	287				
	I_k^{WNC}	0.956	I_{ik}^C 0.273	I_{ik}^C 0.040	I_{ik}^C 0.014				
			I_{ik}^L 0.600	I_{ik}^L 0.100	I_{ik}^L 0.667				

In order to test the impact of the nature of the corpus to index values the validation procedure was performed on a small literary corpus of 0.5 MW for aselected number of literals. The results obtained show that the index values can be largely affected by the nature of the corpus. Thus, for example, the values of both I_{ik}^C and I_{ik}^{WN} have dramatically changed for senses 1a and 2a of the noun *lice* (Table 3). This does not come as too much of a surprise

since meaning 2a ("A part of a person that is used to refer to a person") is more used in newspaper texts whereas the meaning 1a ("The front of the human head...") in literature. The changes seem to be far less dramatic for the indices I_{ik}^{L}, but in order to draw some final conclusions the literals should be tested on a larger corpus.

4 Conclusion

The applied procedure confirmed the importance of the validation of synsets on a corpus. The adequacy of placement of each literal and its sense in a synset can not be fully assessed without analyzing its appearances in the concordance lines. The frequency indices can serve as useful numerical indicators in this assessment procedure. However, to get a fair estimate of a literal in terms of these parameters, the procedure needs to be applied on a large and balanced corpus. To that end automatic or/and semi-automatic procedures need to be developed in order to alleviate the time-consuming task of manual concordance analysis.

References

1. Kilgarriff, A. and Rosenzweig, J.: English SENSEVAL: Report and Results. In: *Proc. of LREC*, Athens, May–June (2000).
2. Krstev, C., et al.: Corpora Issues in Validation of Serbian Wordnet. In: *Proc. of the Conference "Text, Speech, and Dialogue"*, Springer LNCS, 138–145 (2003).
3. Miller, G. A., et al.: Using a semantic concordance for sense identification. In: *Proc. of the ARPA Human Language Technology Workshop*, 240–243 (1994).
4. Obradović, I., et al.: Application of Intex in Refinement and Validation of Serbian Wordnet. *6th Intex Workshop*, 28–30th May, Sofia (2003).
5. Stamou, S., et al.: BALKANET: A Multilingual Semantic Network for Balkan Languages, *Proc of 1st Global WordNet Conference*, Mysore, India (2002).