Spink, A., Bateman, J. & Jansen, B. (1999). Searching the Web: a survey of EXCITE users. Internet Research: Electronic Networking Applications and Policy, 9 (2), 117-128.

Spink, A., Wolfram, D., Jansen, B. & Saracevic, T. (2001). Searching the Web: the public and their queries. Journal of the American Society for Information Science and Technology, 52 (3), 226-234.

Taniar, D., Jiang, Y., Rahaya, J. & Bishay, L. (2000). Structured Web pages management for efficient data retrieval. Proc. of 1st Int'l Conference on Web Information Systems Engineering (WISE'00), 2097-2104.

Turnbull, D. (2003). Augmenting information seeking on the World Wide Web using collaborative filtering techniques. http://donturn.fis.utoronto.ca/research/augmentis-toc.html (accessed June 2003).

Authors' Information

Peretz Shoval – Department of Information Systems Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: shoval@bgumail.bgu.ac.il

Tzvi Kuflik – Department of Management Information Systems, Haifa University, Haifa, Israel;
e-mail: tsvikak@mis.hevra.haifa.ac.il

EMPIRICAL METHODS FOR DEVELOPMENT AND EXPANDING OF THE BULGARIAN WORDNET

Pavlina Ivanova, George Totkov, and Tatiana Kalcheva

Abstract: Some basic points from the automated creation of a Bulgarian WordNet – an analogue of the Princeton WordNet, are treated. The used computer tools, the received results and their estimation are discussed. A side effect from the proposed approach is the receiving of patterns for the Bulgarian syntactic analyzer.

Keywords: Empirical Methods in NLP, WordNet

1. Introduction

WordNet is developed in the Princeton University [2,4] as a lexical database of English. The first multilingual database to realize such approach is *EuroWordNet* (EWN) ([11], [12]) consisting of eight European languages. The monolingual databases are related to the *Princeton WordNet* (PWN) (and in this way to each other) via an interlingual index (ILI).

The *Bulgarian WN* (BWN) has been developed as a cooperative task involving the Plovdiv University and the Department for Computer Modelling of Bulgarian Language at the Bulgarian Academy of Sciences (DCMB). The work is part of an EC funded project (IST-2000-29388) *BalkaNet* [7] for the creation of a multilingual lexical database (like EWN) for 6 Balkan languages (Bulgarian, Greek, Romanian, Serbian and Turkish, Czech).

2. Forming of a BWN

The main stages in the automatic creation of a BWN (A_BWN) are presented in [8]. We discuss further the tools and the results received in this process – namely the extraction of synsets from an English-Bulgarian dictionary (EBD) and the receiving of A_BWN.

Our starting point is the transformed EBD [6] with more than 160,000 entries. Each different meaning of an English word is placed on a different row. Each row contains the English word (entry) and its translation equivalents (TE) in Bulgarian. A link is added (where it was possible) between the EBD rows and the PWN synsets (via the ILI) [9].

Each TE row may contain Bulgarian words and phrases separated with the following signs: comma, colon, semi-colon, full stop, slash and brackets. In order to receive the different synonyms from a TE row we had to differentiate the punctuation marks used as 'separators' from the ones marking some orthographical rule. E.g. in the translation of "*anticipant*" – *"човек, който чака, чакащ"*, the first comma is not a separator while the second one is.

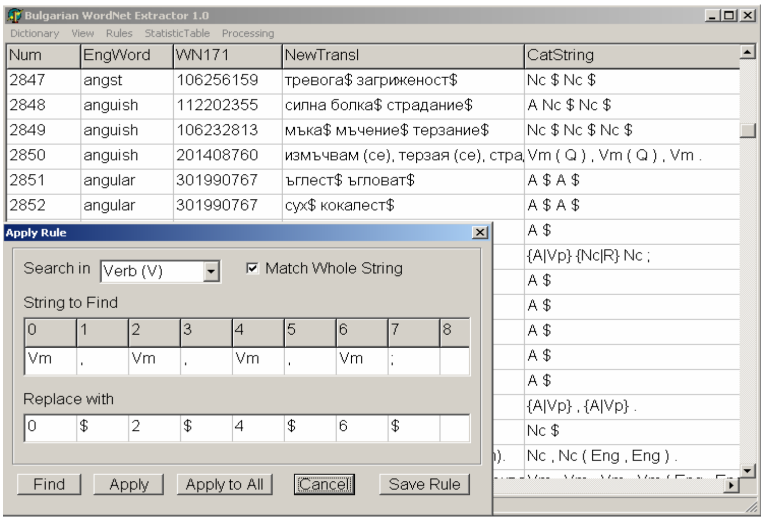
A special tool *BWN Extractor* (BWNE) is designed for the solving of the problem. The BWNE was created to extract almost automatically meaningful rules for forming Bulgarian synsets corresponding to PWN. In the first place, the Bulgarian words in TE rows were processed by Bulgarian Morphological Analyzer BulMorph 2.0 [10] in order to get a list of their morphological characteristics (MC). As a result we received a string-pattern in which every Bulgarian word from the TE row was replaced with a special symbol(s) coding its MCs (e. g. N denotes a noun, A – adjective, V – verb, D – adverb, Vm – a verb in indicative mood, Va – the verb 'be', Vp – participle, Nc – common noun, Q – particle, etc.) The morphological alternatives (ambiguities) are separated with '|' and the results from the robust morphological analysis [10] are marked with the sign '^'.

Table 1 presents syntactic patterns (SynP), obtained with BWNE and ordered according to their frequency in the processed TE.

SynP	Noun	Verb	Adjective	Adverb	Total
Nc	10134	11	45	2	10192
Nc , Nc	4123	5	8	0	4136
A	59	2	3749	25	3835
Vm	20	2463	24	1	2508
A , A	13	0	2460	11	2484
Vm , Vm	11	2215	8	2	2236
A Nc	2070	2	36	1	2109
Nc , Nc , Nc	1009	0	2	0	1011
Nc|Vn	913	0	5	0	918

Table 1. The first 9 syntactic patterns received by BWNE

What this statistics shows is that, for example, when the TE row of an English word consists of two nouns separated by comma (Nc, Nc) in 4123 of 4136 cases (more than 99.6%) the English word is also a noun and the corresponding two Bulgarian words (nouns) are two synonyms. Only the cases when the part of speech (POS) does not match are questionable and need to be marked by expert using BWNE.


Bulgarian WordNet Extractor 1.0
Dictionary View Rules StatisticTable Processing
Apply Rule
Find
Apply
Apply to All
Cancel
Save Rule

Num	EngWord	WN171	NewTransl	CatString
2847	angst	106256159	тревога$ загриженост$	Nc $ Nc $
2848	anguish	112202355	силна болка$ страдание$	A Nc $ Nc $
2849	anguish	106232813	мъка$ мъчение$ терзание$	Nc $ Nc $ Nc $
2850	anguish	201408760	измъчвам (се), терзая (се), стра	Vm (Q) , Vm (Q) , Vm .
2851	angular	301990767	ъглест$ ъгловат$	A $ A $
2852	angular	301990767	сух$ кокалест$	A $ A $
				A $
				{A|Vp} {Nc|R} Nc ;
				A $
				A $
				A $
				A $
				A $
				{A|Vp} , {A|Vp} .
				Nc $
			).	Nc , Nc (Eng , Eng) .

Search in Verb (V) ☑ Match Whole String

String to Find

0	1	2	3	4	5	6	7	8
Vm	,	Vm	,	Vm	,	Vm	;	

Replace with

0	$	2	$	4	$	6	$	

Figure 1. The Rule Editor window of the BWNE

The rules for the separation of the synonyms are based on the automatically received SynP. Moreover, BWNE provides a special *Rule Editor*. Figure 1 shows the creation of a rule to be applied on all rows corresponding to an English entry defined as 'verb'.

The functional capabilities of the *Rule Editor* are: a) automatically synthesizes rules, starting with the most likely ones; b) allows additional editing of the automatically synthesised rules; c) represents all the rows in EBD corresponding to the processed SynPs in *View* mode; d) allows changes in the respective rows in EBD in *Edit* mode; e) gives possibility for successive processing of rows from EBD (one by one or in group) in *Apply* mode; f) provides *Save Rule* mode, etc.

Experiments show that approximately 45,000 rows (TE) from the initial EBD can be automatically processed with the first 100 synthesized rules. The next 3,000 rules process additional 20,000 rows. In this way about 65,000 rows of EBD are almost automatically processed with 3,300 rules. The extracted synonyms form A_BWN, containing about 42,000 Bulgarian synsets linked to the corresponding English synsets in PWN.

Table 2 presents 15 of the 3,300 automatically synthesized rules. Each rule consists of three parts: *POS* of the entry for whose TE a given rule is applied; *Left side* containing the (searched) string-pattern and *Right side* defining the replace string – a sequence of numbers (position of the Left side components) separated by the sign '$'. E.g. rule 15 means that 4 synonyms will be extracted in all the TE rows (for which the ILI corresponds to a 'verb') matching the pattern ***verb1/verb2 noun1/noun2.*** The four extracted synonyms (separated by '$') are as follows: ***verb1 noun1 $ verb1 noun2 $ verb2 noun1 $ verb2 noun2$.***

Note that in rules 9-12 the comma is not (always) a separator. Its role depends from the POS of the entry – a comma followed by a relative pronoun (Pr) is a separator when the corresponding POS is A (rule 1) but it isn't when the POS is N (rule 10).

№	POS	Left Side	Right Side
1.	A	A , A , Pr Pp Vm	1 $ 3 $ 5 6 7 $
2.	A	D {A|Vp} , A	1 2 $ 4 $
3.	A	A ; R A Nc	1 $ 3 4 5 $
4.	A	Vp , A , A , R A Nc	1 $ 3 $ 5 $ 7 8 9 $
5.	D	D , D , R Pd {A|Nc}	1 $ 3 $ 5 6 7 $
6.	D	R A Nc / Nc	1 2 3 $ 1 2 5 $
7.	N	A / Vp Nc	1 4 $ 3 4 $
8.	N	A Nc , {An|D} Nc , {An|D|Nc}^	1 2 $ 4 5 $ 7 $
9.	N	An Nc , Vp R A Nc	1 2 3 4 5 6 7 $
10.	N	Nc , Pr Vm / Vm	1 2 3 4 $ 1 2 3 6 $
11.	N	Nc , R Pr Q Vm Nc	1 2 3 4 5 6 7 $
12.	V	Vm (Nc , Nc , {Nc|Np}) ; Vm	1 2 3 4 5 6 7 8 $ 10 $
13.	V	Vm (Q) , Vm (D)	1 3 $ 1 $ 6 7 8 9 $
14.	V	Vm , Nc Va R	1 $ 3 4 5 $
15.	V	Vm / Vm Nc / Nc	1 4 $ 1 6 $ 3 4 $ 3 6 $

Table 2. Rules for the extraction of Bulgarian synonyms

3. Evaluation of the A_BWN

In order to validate the A_BWN we used BWN prototype[1]. The presented result is for an A_BWN consisting of 39,109 Bulgarian synsets and containing 9,936 (common) ILI with the BWN prototype.

Let denote the number of the common literals (different words and phrases in a synset) with E, the number of the A_BWN literals –with F and the number of the A_BWN literals in the intersection – with P[1]. In order to estimate the A_BWN we use two measures:

[1] The prototype, containing 15,007 Bulgarian synsets, is created (manually) by the DCMB experts.

$$\text{Precision} = \frac{P}{F} \text{ and } \text{Recall} = \frac{P}{E}.$$

The number of literals in the BWN prototype is 18,520 and in the A_BWN – 21,302. The average number of literals in a synset is 1.864 and 2.144 respectively. The number of literals common to A_BWN and the BWN prototype is 9,449. The number of synsets common to A_BWN and the BWN is 9,936. The average number of common literals in a synset is 0.951. The *Recall* is 51.02% and the *Precision* is 44.36%.

The new synsets in A_BWN (more than 33,000 additional ILI) give opportunity for further expanding of the BWN prototype.

4. Receiving of Syntactic Patterns

A side effect of the proposed approach is the receiving of syntactic patterns for 4 phrase types in Bulgarian: NP (noun phrase), VP (verb phrase), AP (adjective phrase) and AdvP (adverbial phrase). For example Table 3 presents the first 4 (applied) rules for A (see Table 2).

№	POS	Right Side
1.	A	A $ A $ Pr Pp Vm $
2.	A	D {A|Vp} $ A $
3.	A	A $ R A Nc$
4.	A	Vp $ A $ A $ R A Nc$

Table 3. The (applied) rules 1-4 from Table 2

In fact the received SP for the structure of AP in Bulgarian:

AP := A | Pr Pp Vm | D {A|Vp} | Pr Q Vm | R A Nc | Vp

has to be checked by expert.

The first 10 SP (with greatest frequency) are presented in Table 4.

The experiments show that in this way we define some meaningful rules for the structure of NP, VP, AP and AdvP. The most frequent patterns are most likely to produce correct rules. Using the proposed approach we received 1762 syntactic patterns for the Bulgarian phrases: 1470 for NP, 175 – AP, 169 – VP and 79 – AdvP.

№	SyntacticPattern	NP	VP	AP	AdvP	Total
1.	Nc	**10744**	3	26	2	10775
2.	A	57	0	**3786**	14	3857
3.	A Nc	**3553**	1	0	3	3557
4.	Vm	10	3435	34	1	3480
5.	{Nc|Vn}	**1328**	4	3	0	1335
6.	Vm Q	0	**958**	2	0	960
7.	Vp	39	0	**887**	7	933
8.	Nc^	**871**	1	1	0	873
9.	Vm R Nc	1	**782**	0	0	783
10.	Vm Nc	2	**725**	0	0	727
Total		26028	8187	7336	902	42453

Table 4. The first 10 syntactic patterns

[1] The literals that don't match the literals in the BWN prototype are not necessarily "incorrect".

5. Perspectives

A method for improvement of Bulgarian Synonym Dictionary (BDS) and removing logical discrepancies in synonym rows is described in [3, 9]. The next step to be done is the expanding and correction of the synsets in A_BWN using the improved synsets from regular BDS [5].

A tool analogous to the *Split/Merge* program [9] is under development. The main features of the tool are: a) displaying all the synsets from A_BWN and BDS, in which a chosen word (or phrase) takes part; b) choice of an A_BWN synset to be processed; c) finding the BDS rows which are closest to the chosen synset [9].

The method for extracting syntactic patterns can be applied to *other lexical resources*, for example to Bulgarian Thesaurus [1]. Additional MCs (number, gender, definiteness, etc.) can be used for synthesis of more precise syntactic rules.

The receiving of precise syntactic patterns can be used for the almost automatic creation of a *Bulgarian computer grammar* (including thousands of syntactic rules). The creation of the computer grammar is a crucial step towards the development of a syntactic analyzer of Bulgarian texts.

References

1. Andrejchin L. (ed.), Bulgarian Explanatory Dictionary. Sofia, Nauka i Izkustvo, 1999 (in Bulgarian).
2. Fellbaum C. (ed.), WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, London, England, 1998.
3. Ivanova P., Totkov G., Automated Improving and Forming Synsets on Conventional (non computer based) Synonym Dictionaries, Proceedings of the International Conf. Automation and informatics'2002, Sofia, 33-36.
4. Miller G., R.Beckwith, C. Fellbaum, D. Gross and K.Miller, Introduction to WordNet: an on-line lexical database. In: International Journal of Lexicography 3(4), 1993, accessible at ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps.
5. Nanov L, A. Nanova, Bulgarian Synonym Dictionary. Sofia, Hejzal, 2000 (in Bulgarian).
6. Rankova M., T. Atanasova, I. Harlakova. English-Bulgarian Dictionary. Izd. Nauka I izkustvo, Sofia, 1990.
7. Stamou S., K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufiş, S. Koeva, G. Totkov, D. Dutoit, M. Grigoriadou, BALKANET: A Multilingual Semantic Network for the Balkan Languages, Proceedings of the International Wordnet Conference, Mysore, India, 21-25 January 2002, 12-14.
8. Totkov G., Towards Building Bulgarian WordNet: Language Resources and Tools, Proceedings of the ICT&P'03, Sofia, 2003, 31-40.
9. Totkov G., P. Ivanova, Iv. Riskov, Automated Improving and Forming WordNet Synsets on Conventional (non computer based) Synonym and Bilingual Dictionaries. in A. Narin'iyani (ed.), Comp. Ling. and its Applications, Proc. of the Int. Workshop DIALOGUE'2003, Protvino, June 2003.
10. Totkov G., R. Doneva, Bipartite Finite State Transducers as Morphology Analyser, Synthesizer, Lemmatizer and Unknown-Word Guesser, Proc. of 2nd Intern. Seminar „Computer Treatment of Slavonic Languages" SLOVKO'2003, Oct. 24-25, 2003, Bratislava (in print).
11. Vossen P. (ed.), EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Final Document, 1998, 108p.
12. Vossen P. Building a multilingual database with wordnets for several European languages. http://www.hum.uva.nl/~ewn, 1999.

Authors' Information

Pavlina Ivanova – pavlina@pu.acad.bg

George Totkov – totkov@pu.acad.bg

Tatiana Kalcheva – selinashery@abv.bg

Plovdiv University, 4 Tzar Asen str., 4000 Plovdiv, Bulgaria.