Semi-Automated English-Russian WordNet Construction: Initial Resources, Software and Methods of Translation

Sergey Yablonsky[1,2]
[1]Petersburg Transport University
Moscow av., 9, St.-Petersburg, 190031, Russia
[2]Russicon Company,
Kazanskaya str., 56, ap.2, 190000, Russia
serge_yablonsky@hotmail.com

Andrey Sukhonogov
Petersburg Transport University
Moscow av., 9, St.-Petersburg
190031, Russia
ASukhonogov@rambler.ru

Abstract

The idea of Princeton WordNet (PWN) transformation into multilingual lexical ontology has started to be put into practice in EuroWordNet project. For today exists more than 15 national versions of WordNet, and all of them are to some extent adhered to PWN. Conformity is reached or by means of interlingual indexes development, or as such index acts PWN. The purpose of the present work is research and development of semi automated methods of English-Russian version of WordNet database (English-Russian WordNet – ERWN) construction using mapping of PWN to RWN and preliminary test translation of PWN/RWN for an estimation of an opportunity of such mapping construction on the basis of PWN. It is shown that up to 70% PWN synsets could be translated into Russian on the basis of the semi-automated translation methods.

Introduction

The idea of Princeton WordNet (PWN) (Miller G. et al., 1990, Fellbaum C., 1998) transformation into multilingual lexical ontology has started to be put into practice in EuroWordNet project. For today there exists more than 15 national versions of WordNet, and all of them are to some extent adhered to PWN. Conformity is reached or by means of interlingual indexes development, or as such index acts PWN.

The purpose of the work is:

• Research and development of methodsgf the semi automated construction of English-Russian version WordNet;

• greliminary test translation of PWN/RWN.

1 Initial resources and software

Initial resources:

- Frequency lexicon from Yandex (http://www.yandex.ru) – 65 000 lemmas from 1500000 search inquiry's word forms.
- English-Russian and Russian-English Oxford Russian Dictionary, 3dgdition. With the purpose of increase of completeness of a covering the Russian-English and English-Russian dictionaries distributed under license GPL in structure of a server of dictionaries DICT Server - RFC2229, and a number of other dictionaries, freely distributed on the Internet are used.
- Lexical DB Princeton WordNet 2.0 (PWN) and Russian WordNet (RWN) (Balkova V., Sukhonogov A.M., Yablonsky S.A. 2004).

All programs are developed on the basis of technology Oracle/J2EE and include:

- English-Russian and Russian-English dictionaries (ER-dictionary) database and API (Application Programming Interface), allowing search and editing of words translations of words. Generally, the dictionary contains translations of a kindg $\{w_1,w_2,..w_n\}$ – $\{t_1,t_2,..t_m\}$, i.e. conformity of set of words of one language to set of words of another;
  - API DB PWN/RWN, allowing to carry out search and editing English-Russian WN.
- Russicongorphological analyzer / normalyzer for Russian language (>160 thousand lemmas) (Yablonsky S.A. 1998, 2004).
- A package of special utilities for various transformations of initial dictionaries, databases, lexicons and so forth.
- Editor TenDrow for editing WordNet.gow editor TenDrow:
  - Works with СУБД Oracle9i/10g and Interbase/Firebird;
  - Carries out data exchange between a DB and OWL-representation;
  - Supports formats of lexical files Princeton WordNet 2.1 and VisDic1.3.36 (for loading in a DB).

2 Methods of translation

Two complementary approaches were devised in EuroWordNet (Vossen, P., 1998) and in BalkaNet (BalkaNet, 2004) to build local wordnets from scratch:

The merge approach: building taxonomies from monolingual lexical resources and then, making a mapping process using bilingual dictionaries.

The expand approach: mapping directly local words to English synsets using bilingual dictionaries.

The merge approach is present in our Russian WordNet (Balkova V., Sukhonogov A. M., Yablonsky S. A., 2004) construction process from the beginning.

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 345–347.

At the same time we use the expand approach for direct mapping of many words from PWN to Russian and vise versa. This approach is also used for English and Russian proper and geographical names.

We engaged the initial hypothesis that nominal hierarchies in English and Russian should be similar, at least for basic domains. This enabled us to formulate our first stage task to attaching RWN nominal entries using English-Russian bilingual dictionary to PWN 2.1 synsets. Like this, the English nominal hierarchy of WordNet serves as a skeleton structure to support the construction of the core Russian nominal WordNet. This approach was also taken up in construction of the Spanish and Catalan WordNets (Atserias J., Climent S., Farreres X., Rigau G. and Rodriguez H., 1997), Italian MultiWordNet (Magnini B., Cavaglia B., 2000; Pianta E., Bentivogli L., Christian Girardi, 2002) and Hungarian WordNet (Márton Miháltz and Gábor Prószéky, 2004).

3 Features of the WordNet translation

Let consider that wordW_{L2} is the translation of word W_{L1} for two languages $L1$ and $L2$ and words have the same POS.

Let consider that following conditions are satisfied for each pair <W$_{L2}$,W$_{L1}$>:

- WordNet for language L1g (WN$_{L1}$)g contains lemm(W$_{L1}$) and WordNet for language L2 (W$_{L2}$) contains lemm(W$_{L2}$), where function lemm (W) will transform word form W to a lemma;
- All possible values lemmg(W$_{L1}$)g are present in gWN$_{L1}$g and all possible values lemm(W$_{L1}$) gare present in$_{g}$ WN$_{L2}$;
- WN$_{L1}$g and$_{g}$ WN$_{L2}$g are connected through mapping WN$_{L1}$(PWN) to WN$_{L2}$(RWN) where it is possible. Here PWN acts as an ILI. In some cases we need to make vise versa mapping of WN$_{L2}$(RWN) to WN$_{L1}$(PWN) if some words/synsets are absent in PWN (mostly Russian geographical and proper names).

Then it could be possible to connect pair <WN$_{L1}$,WN$_{L2}$> for languages L1 and L2 so, that the algorithm of connection will connect all synsetsWN$_{L1}$ to all synsetsWN$_{L2}$ are using WN$_{L1}$(PWN) as an ILI.

Generally such display is impracticable, as:

- For some wordW$_{L1}$ can not exist corresponding wordW$_{L2}$, i.e. translation can be absent;
- The number of values lemm (W$_{L1}$) can be not equal to number of values lemm(W$_{L2}$) and/or values can not coincide;
- Some wordW$_{L1}$ can be translated not in a wordW$_{L2}$, but in some word-combination which is not a phrase unit in language L2 and vise versa.

Therefore ideal translation WordNet from one language on another (in our case from English on Russian and vise versa) does not exist, and the algorithm of construction of mapping of PWN to RWN should resolve above contradictions.

For reception of correct conformity PWN and RWN synsets it is necessary to have correct translation of one of synset's words.

According to the accepted order of construction of an index, initial is PWN synset - S_{PWN}.

At the first stage for this synset the list of alternatives synsets from RWN – S_{RWN} = { S$_{RWN1}$, S$_{RWN2}$, …, S$_{RWNn}$ } is under construction.

IfS_{RWN}={Ø}, two variants exist:

- There are translations of S_{PWN} but in RWN there are no corresponding synsets. In this case it is possible to create new synsets from translations.
- If translations are absent, construction of an index for S_{PWN} is impossible.

Both variants demand additional manual processing.

IfS_{RWN} ≠{Ø}, for each synset value of estimated function$R = R(\mathrm{S}_{PWN}, \mathrm{S}_{RWNi})$ is calculated.

At the second stage for every synsetS$_{RWNi}$ fromS_{RWN} the list of synsets-alternatives PWN -S $_{PWN}$ = { S$_{PWN1}$, S$_{PWN2...,}$S$_{PWNm}$} is formed and the value of estimated function$R'_i = R'(\mathrm{S}_{RWNi}, \mathrm{S}_{PWN})$ is calculated. In the list$S_{PWNi}$of every synset-alternative fromS$_{PWN}$ the initial synsetS$_{PWN}$ is always putted because the uniform English-Russian and Russian-English dictionary is used.

4 General algorithm of PWN-RWN translation

Let S_{PWN} - *PWN* synset,$S_{RWN.}$– *RWN* synset.

S_{PWN} = *{w1, w2 … wn}, wi* - a lexeme of $S_{PWN.}$ synset.

$T(w)$ - a set of translations of a lemma$w$ in the generalized ER-dictionary.

The whole process of translation could be divided in such main steps.

1) By means of the ER-dictionary we form translations: $T(S_{PWN.})$ = *{*$T(_{w1})$, $T(_{w2})$*, …* $T(_{wn})$*}*

It is possible to present $T(S_{PWN})$ in the form of a matrix with rows - lexemes of S$_{PWN}$synset, and columns - variants of translation of a lexeme in the ER-dictionary:

$T(S_{PWN})$ = *{{* $p_{11,} p_{12,} \ldots p_i$*},{* $p_{21,} p_{22,} \ldots p_{2k}$*},…* *{*$p_{n1,} p_{n2,} \ldots p_{nm}$*}}*

where p_{ij} - *j-th* word from the set of translations of a wordw_i.

2) Let *TS (*S$_{PWN.}$) be a set of the most probable translations S$_{PWN.}$, *TS (*S$_{PWN.}$) ?T(S$_{PWN.}$). Then function *DC* is defined as: *TS (*S$_{PWN.}$*) = DC (T (*S$_{PWN.}$)).Function *DC* allocates from$T$(S$_{PWN.}$) the subset of translations consisting of translations$T(w)$ for which power of crossing$m$ = $T(w_{ij}$*)*∩*…* ∩*T(*w_{nm}) is maximal. We are looking for translations containing the maximal number of concurrences of words-translations.

3) If *TS(*S_{PWN}*)*=∅, there are no translations (the words are absent in the RWN lexicon) and these words are excluded from the received set of translations *TS (SPWN).*

4) If $m(TS(S_{PWN}))$ = 1, the ER-dictionary contains unique (monosemic) translation of SPWN synset. Let$T(S_{PWN})$ = $\{ p_{1,} p_{2,} \ldots p_n\}$ is a unique variant of translation$S_{PWN.}$For each word of set $\{p_{1,} p_{2,} \ldots p_n\}$ the set of RWN synsets$[S_{RWN}]_i$ is formed.$R(S_{PWN})$ = $\{[S_{RWN}]_{1,} [S_{RWN}]_{2,} \ldots [S_{RWN}]_n\}$. At construction of set $R(S_{PWN})$ the condition of part of speech conformity for PWN and RWN synsets is checked. The most corresponding translation is taken from the set $R(S_{PWN})$.

5) If $m(TS\ (_{SPWN})) > 1$, the ER-dictionary contains set of variants of lemmasS_{PWN} translations. For each variant the set of translations $R(S_{PWN})$ = $\{[S_{RWN}]_1$, $[S_{RWN}]_{2,} \ldots [S_{RWN}]_n\}$ is formed (the condition of part of speech conformity is checked). At this step we designed several web-oriented methods of automatic or semi-automatic translation based on automatic meaning discovery using Normalized Google Distance (Sukhonogov A., Yablonsky S., 2005) to process multilingual WordNets for Russian.

5 Conclusion

Preliminary test translation and mapping of PWN/RWN has shown that it is possible to translate and bind to Russian WordNet up to 70% of PWN synsets by means of the semi-automated connection between PWN and RWN synsets. In the Table 1 we show the number of not translated Lemms, Lexems and Synsets of PWN after preliminary test semi-automated translation. The main not translated words are proper and geographical names, chemical and medical terminology. We assume that the English synset is translated if one member of the synset is translated (into Russian).

Table 1:

	Noun	Adjective	Verb	Adverb
Lemm	48287	2366	760	855
Lexem	49756	2412	810	888
Synset	28871	1876	642	754

6 Acknowledgements

Our thanks go to our colleague Valentina Balkova that takes definite part in Russian WordNet development.

References

Fellbaum C. (1998) WordNet: an Electronic Lexical Database. MIT Press, Cambridge, MA.

Miller G. et al. (1990) Five Papers on WordNet. CSL-Report, vol.43., Princeton University; `ftp://ftp.cogsci.priceton.edu/pub/wordnet/5papers.ps`.

Vossen, P. (1998) EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht: Kluwer.

BalkaNet (2004) – Design and Development of a Multilingual Balkan WordNet. Romanian Journal of Information Science and Technology Special Issue (volume 7, No. 1-2); `http://www.ceid.upatras.gr/Balkanet/`.

Atserias J., Climent S., Farreres X., Rigau G. and Rodriguez H. (1997) Combining multiple methods for the automatic construction of multINDngual WordNets. Proceedings of the International Conference "Recent Advances on Natural Language Processing" RANLP '97, Tzigov Chark, Bulgaria.

Pianta E., Bentivogli L., Christian Girardi (2002) MultiWordNet: developing an aligned multilingual database." In Proceedings of the First International Conference on Global WordNet, Mysore, India, January 21–25.

Magnini B., Cavaglia B. (2000) Integrating Subject Field Codes into WordNet // Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation / M. GavrINDdou, G. Crayannis, S. Markantonatu, S. Piperidis, G. Stainhaouer (eds.). – Athens, Greece, 31 May – 2 June. – pp. 1413–1418.

Márton Miháltz and Gábor Prószéky (2004) Results and Evaluation of Hungarian Nominal WordNet v1.0. In Proceedings of the Second International WordNet Conference (GWC 2004), Brno, Czech Republic, January 20–23.

Yablonsky S. A. (1998) Russicon Slavonic Language Resources and Software. In: A. Rubio, N. Gallardo, R. Castro & A. Tejada (eds.) Proceedings First International Conference on Language Resources & Evaluation, (pp. 1141–1147). – Granada, Spain.

Yablonsky S. A. (2004) Integration of Russian Language Resources. In: Proceedings 4th International Conference on Language Resources & Evaluation. – Centro Cultural de Belem, Lisbon, Portugal.

Balkova V., Sukhonogov A. M., Yablonsky S. A. (2004) Russian WordNet: From UML-notation to Internet/Intranet Database Implementation. In Proceedings of the Second International WordNet Conference (GWC 2004), Brno, 2003. – P. 31–38.

Sukhonogov A., Yablonsky S. (2005) Semi-automatic English-Russian WordNet translation. In Proceedings of the Dialogue 2005 International Conference on the Computational Linguistics and Intellectual Technologies. Moscow: Nauka (in Russian).