DEVELOPMENT OF LEXICAL BASIS FOR THE UNIVERSAL DICTIONARY OF UNL CONCEPTS

Dikonov V. G. (dikonov@iitp.ru)

IITP RAS, Moscow, Russia

The paper describes the current state of development of the lexical basis of an open and free lexical-semantic resource — the Universal Dictionary of UNL Concepts (UNLDC). The resource serves as a lexicon of an artificial intermediary language UNL (Universal Networking Language). It links the elementary units of UNL — concepts with lexicons of natural languages and various external lexical and semantic resources, including Wordnet and SUMO ontology. The dictionary's main goal is to support automated semantic analysis, encoding the meaning of the text as UNL semantic graphs and subsequent generation of text in different natural languages.

Keywords: lexical resources, semantics, interlingua, UNL

1. Introduction

The dictionary of the artificial interlingua UNL (Universal Networking Language), also called Universal Dictionary of UNL Concepts (UNLDC) is a part of an international project to develop UNL [Boguslavsky, et. Al, 2005], [UNL Specification 2005]. The development of this resource is supported by the members of the "U++ Consortium", which unites researchers from Russia, France, Spain and India. Although UNL is the main application of the resource, it also has certain value of its own and can be used for scientific and practical tasks not directly related to UNL.

The basic units of the dictionary are so called UNL concepts, which correspond to word senses described by traditional explanatory dictionaries. They are also similar to semanthemes in the Meaning↔Text theory by I. A. Melchuk, which includes deep syntax and semantic representations with many features in common with semantic graphs of UNL. The notion of UNL concept is well aligned with lexicographic tradition. This allows to reuse much of the natural language data already gathered by explanatory dictionaries and thesauri.

UNLDC defines the inventory of concepts in U++ UNL. Each concept receives a unique identifier called Universal Word (UW) conforming to the U++ standard. Each new UW should be added to the dictionary before being used in any UNL encoded documents in order to maintain lexical compatibility between different software tools supporting generation of natural language text from UNL. UNLDC links UWs with words and expressions of natural languages that can be used to express corresponding concepts.

The dictionary has three main parts:

1. List of the U++ Universal Words (UWs),
2. Semantic network that links the UWs together,
3. Set of local dictionaries linking UWs with words and expressions of natural languages.

Entries have links to external lexical and semantic resources, including internal dictionaries of several MT systems. The UWs and their links to natural languages are annotated to keep track of their source, status and expected quality. All data is split into several complementing each other "volumes" to simplify maintenance. Each volume is stored in a separate file. Different volumes may be used to represent languages and alternative orthographies/dialects of the same language. Large groups of entries, such as domain terms or named entities, are split away into separate volumes too.

The general overview and full introduction to UNLDC has been given in [Dikonov V., Boguslavsky I., 2009]. In this paper we describe the results of our work done in 2012 to extend the lexical coverage of UNLDC in Russian and integrate data for other natural languages provided by our partners.

2. Current state of the project

At the time of writing the total number of UWs in the dictionary has reached 2,880,661. There are seven local dictionaries of Russian, English, French, Hindi, Spanish, Vietnamese and Malay. A considerable part of the semantic network is completed. The core of the semantic network — its ontological structure — is modeled on the basis of the SUMO ontology [Pease, 2011].

The main priority is given to the core lexical part of UNLDC, which covers most frequently used words of English and Russian, and some special semantic units corresponding to lexical functions, modal words and some closed class words. In addition to this part we develop separate extension volumes containing basic terminology and named entities. The extensions are still at an early stage and contain only automatically gathered data. Table 1 shows, how many UWs are there in each part.

Table 1. Number of UWs by dictionary part in early 2013

Part	UWnumber	File	Status
General lexics	82,804	CommonUNLdict-XML-0.04.1-alpha.tar.bz2	Downloadable
Terminology	688,617	CommonUNLdict-CSV-Terminology-0.02.tar.bz2	Under development
Named Entities	2,109,240	CommonUNLdict-CSV-NamedEntities-0.01.tar.bz2	Soon to be released

The principle approach to the development of the resource is accumulation and integration of data available in the Internet with subsequent proofreading and gap filling. The initial versions of the local dictionaries are built using automated methods of import and cross-linking of different sources. The preferred sources are more reliable ones, such as translation dictionaries, or semantically annotated resources (Wordnets, ontologies, other UNL dictionaries). Since every resource contains some errors and automatic integration tends to multiply them, we have to to put a lot of effort into finding and fixing the resulting "noise". It is planned that all automatically gathered data will eventually be proofread. However, the manual verification process is far too labor and time consuming to be applied to all the data. Therefore, proofreading is done only for the most important parts of the resource. Some errors are detected by applying formal criteria. For example, if an UW is linked to words of several languages, but no translation dictionary confirms that these words are good translation equivalents of each other, we can suspect that some of the word↔UW links are wrong and lower their reliability score. Another way to sift through the data automatically is to check if the semantic argument frame manually ascribed to the UW matches its taxonomic class in the associated ontology and if both are compatible with any external semantic annotations linked to the corresponding words of natural languages.

2.1. Local Dictionaries

The existing local dictionaries are not equal in size and quality. They have been built using different approaches and from different data. The most interesting part is the general lexics part, because it is being proofread by hand. Table 2 quotes the number of UWs with translations by language and dictionary part with rough quality estimation.

Table 2. Number of UWs linked to words and expressions of natural languages in early 2013

Language	General lexics	Terminology	Names	Total	Quality estimation
English	**82,804** (40,894 words)	688,617	2,109,240	2,880,661	*****
Russian	**48,555** (30,818 words)	688,613	226,595	963,763	**** Manual proofreading in progress
French	**36,324** (25,068 words)	103,060	367,888	507,272	*** Automatic verification
Hindi	**27,815** (30,220 words)	0	10,823	38,638	*** Automatic verification
Spanish	**11,758** (6,983 words)	21,990	298,674	332,422	** Experimental
Malay	**21,861** (17,457 words)	0	46,044	67,905	** Experimental
Vietnamese	**5,927** (6,456 words)	0	171,367	177,294	*** Experimental

2.2. Volumes of general lexics

The English general lexics volume was built from the Princeton Wordnet 2.1 as a result of the work done by the Spanish UNL center. English is used to form the majority of UWs, so almost all UWs links to some English words or phrases. The Wordnet data were supplemented with new UWs standing for semantically non-void prepositions and conjunctions and some phrasal verbs. 13,811 UWs representing lexical senses of 6,723 English words were rewritten or created by hand. Those changes concerned the most frequent English words. Further, manual changes were made in about 5,000 predicate UWs to fix bad argument frame descriptions. The general English local dictionary does not contain all of the Wordnet. We rejected all multiword expressions from Wordnet and chose about 40,000 most frequent words to facilitate linking to the internal dictionary of the ETAP-3 MT system.

The Russian general volume 0.05-alpha registers over 48,000 senses of 30,800 Russian words. It is being gradually improved by the author through proofreading and adding new data. It still includes only those Russian words and word senses (with very few exceptions) that serve as translations of already existing English Wordnet senses. This happens due to the nature of the process used to construct the initial version of the Russian dictionary and the necessity to a) clean up the unavoidable errors before adding more specifically Russian lexical data and b) to maximize the percentage of UWs that have translations into all supported natural languages. The next version 0.06 is going to include Russian words that lack direct single word translations into English. The examples include such common words as *старик (old man), касса (cash register), телевизор (TV set), молчать (keep silence), белеть (*be seen as white), приходить (come on foot),* historic and cultural phenomena, e.g. *самовар (samovar), щи (cabbage soup)*, *лапоть (peasants' shoe)*, *хохлома (khokhloma)*, *по-черному (house without a chimney), etc*.

The French volume has been built automatically on the basis of the free French Wordnet WOLF 0.1.5 [Sagot, 2008]. It includes only those French words that were linked to the Princeton Wordnet synsets, the members of which had matching UWs in the subset forming the core of UNLDC. The WOLF data were supplemented by the lexical data provided by the French UNL center and further ranked through automatic comparison of the resulting possible French-Russian translations with regular French-Russian dictionaries.

The Hindi volume is made from the UNL dictionary supplied by the Center For Indian Language Technology at the Indian Institute of Technology in Mumbai. It can be expanded using the existing open Hindi Wordnet.

The Spanish volume is based on the small public Spanish Wordnet.

The Malay and Vietnamese volumes are the results of experiments in assimilating lexical data from regular translation dictionaries. We developed tools that can automatically pair the already registered UNL concepts with words/symbols or expressions of arbitrary natural language taken from dictionaries that translate them into the already supported natural languages, e.g. English, Russian and French.

2.3. Links to external resources

UNLDC is being built using data from other open resources and can in turn become a source of data for other projects. The links with external resources are important to enable easy exchange of linguistic data. The UWs in UNLDC are connected with Princeton Wordnet 2.1 and 3.0, Suggested Upper Merged Ontology (SUMO) [Pease, 2011], DbPedia Ontology. Table 3 shows the number of existing links. More external resources may be added to this list in future.

Table 3. Statistics of the links to external resources

Part	UW number	Connected with
General	77,671 77,293	Princeton Wordnet 2.1 и 3.0 SUMO ontology
Terminology	all part	Upper SUMO classes (not reliable) Domain ontologies
Named entities	all	DbPedia ontology Upper SUMO classes

UWs and Russian, English, French and Hindi words also have pointers to the entries of internal dictionaries of linguistic processors supporting those languages. Such connections are needed to convert UNL semantic graphs to text in different languages. The Russian local dictionary links 40,175 UWs with 26,188 entries of the Russian combinatorial dictionary used by the ETAP-3 system. Figure 1 shows the diagram of links between parts of UNLDC and external resources.

The ontology pictured in Fig. 1 is our custom OWL rendering of SUMO. It does not include complex axioms from SUMO due to the differences in expressivity between KIF and OWL languages. The method of producing the OWL version permits periodic synchronization with SUMO and custom changes in the resulting ontology. The ontology is used to form the ontological part of the semantic network connecting all UNL concepts. This part is under active development.

2.4. Data files

The data files are public and can be downloaded under GPLv3+ and Creative Commons CC-BY-SA. Currently the data is provided in two formats: simple tab separated tables CSV and XML for uploading into the lexical database Jibiki/Pivax [Boitet et al., 2007], which can be used as an online search tool. It is also planned to add the RDF/Turtle format conforming to the Semantic Web Linked Data principle. The released data files are available at http://atoum.imag.fr/geta/User/services/pivax/data/. The currently available files include two out of three main parts of the dictionary: the list of concepts (unlvolume) and loacal dictionaries (rusvolume, fravolume, engvolume etc.).

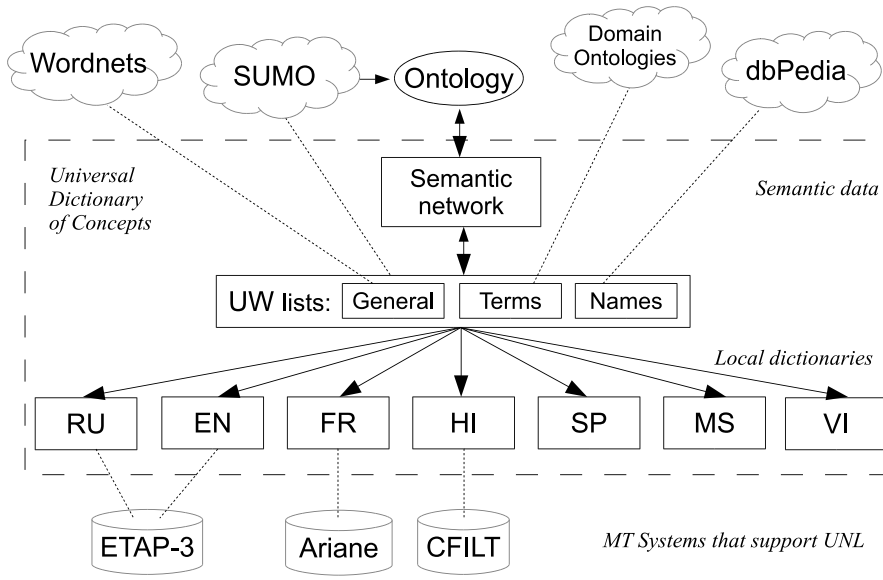

Fig. 1. Links between parts of UNLDC and external resources

3. Some features

Each UW in UNLDC has a special mark showing which language motivated its creation. If it is discovered that translations of the same UW into different languages display subtle differences in meaning or connotations and the conflict is not resolved by ontology, or the ontological classification itself is doubtful, the specified language becomes the final reference.

3.1. Levels of semantic affinity

The dictionary structure allows to distinguish full and quasi synonyms. A single UW may have several translations into the same natural language that are supposed to be full synonyms. For example, the UW *hydroplane(icl>airplane>thing,equ>seaplane)* is linked to two Russian words *гидроплан* and *гидросамолет* with exactly the same meaning. If two words are not interchangeable in most contexts, they should be connected to different UWs. The UWs themselves should than be linked by a synonymy link "equ". Most of the existing synonymy links were imported from Princeton Wordnet. Unlike the Wordnet, groups of synonyms are not treated as units representing one common meaning. Synset members are considered quasisynonyms, unless the opposite is proven. For example, the UWs *hydroplane(icl>airplane>thing,equ>seaplane)* and *seaplane(icl>airplane>thing)* can be merged together, if it turns out that in all languages but English they receive identical translations.

3.2. Special UWs

Usually in order to translate a concept into a natural language one only needs to check, what words are linked to the UW in the local dictionary. However, there are special categories of UWs, which are hard or impossible to translate in the regular way. First of all, it concerns the UWs standing for abstract concepts known as Lexical Functions (LF). The dictionary contains some UWs equivalent to "collocate" type LFs. They may be used to avoid faulty literal interpretation of certain idiomatically used words. For example, the current version of our UNL semantic analyzer renders words *take* and *have* in *I took a short walk* and *I had a short rest* as a special UW *perform_an_action(icl>do,agt>thing,obj>process)*, which corresponds to the abstract meaning of the lexical function OPER1. Its translation into other languages depends on the LF's argument and may be empty. Compare Russian translations *Я немного прогулялся (I walked a little)* and *Я слегка отдохнул (I rested a little)*.

Another group of UWs that are hard to correlate with individual words of natural languages is modal predicates. These UWs are parallel to UNL modal attributes used to encode modality and constitute a well organized system described in [Dikonov, 2009]. In many languages, including English and Russian, words used to express modality are polysemic and represent different modal meanings in different contexts. In UNL the same modal attribute or UW is used to encode a given modality regardless of what modal word was used in the source sentence. E.g. the prohibition in *You may not carry a weapon here* and *You can not smoke onboard* is expressed by two different modal verbs but in UNL the same symbol is used to represent both. The dictionary will link the special modal UW *grant-not(icl>modal>be,obj>uw,aoj>thing)* to both *can not* and *may not* leaving the choice between them to the processor or the user.

3.3. Multiword expressions

Concepts may be translated into some languages by multiword phrases and/or grammar constructions with a variable part. The target language expression may be written not only in the form of a simple n-gram, but also include syntactic relations between the words following the notation of the MT system that is supposed to support the target language. For example, the UW *bathroom(icl>room>thing)* is translated into French as the phrase *salle de bain*, which has the structure recorded in the format of the French MT system Ariane as: *@@_1:'salle' (2:'bain')::1 ": CAT(CATN),GNR(FEM),N(NC). 2 ":CAT(CATN),GNR(MAS), N(NC), NUM(PLU), ART(ABS), RSUNL(MOD)*.

In addition to recording the phrase structure in the formats understood by specific MT systems, UNLDC may use a more general format similar to the text form of UNL graphs. In such case the UWs are replaced with corresponding words of the target natural language. The words are linked with UNL semantic relations and carry UNL attributes. This option is already used in the terminological part of the dictionary. A UNL-capable system should be able to convert such phrase into a fragment of its own internal representation, if applicable. For example, the Russian verb *белеть* in *Белеет парус одинокий* must be translated into English by grammatic constructions

with a variable: *A white* (lonely sail) *can be seen/There is a white* (lonely sail)*/A white* (lonely sail) *appears* (at …). Here the brackets contain a variable NP which cannot be omitted. The English local dictionary might contain the following record:

*mod(*concrete_thing(icl>thing).*@indef.@topic,white),*

obj(see.@entry.@ability, concrete_thing(icl>thing).*@indef.@topic),*

which corresponds to *a white … can be seen*. The UWs in bold instead of words describe the general class uniting the words that could be inserted into the specified slots to build a good translation.

4. Comparable resources

The most important resources similar to UNLDC are the Wordnet family with "Inter Language Index" as a whole and parallel projects building dictionaries for other flavors of UNL. There are two major alternative UNL resources: the dictionary of UNL Development Center developed under the lead of Hiroshi Uchida and UNLarium dictionaries. All of them collect multilingual lexical data and provide some semantic annotation.

4.1. Other UNL dictionaries

The UNL Development Center (UNDC) dictionary, http://www.undl.org/unlexp/ was built by automatic corpus mining using methods common among the developers of statistical and example-based MT systems. This is confirmed by the existence of a large number of characteristic errors. The method does not take into account morphological features and in the case of languages with rich morphology, such as Russian, that dictionary resorts to crude approximation techniques trying to discover lemmas. As a result, the UNDC dictionary contains a lot of wrong lemmas and false translations. There are some differences in the UW formation standards too. UNDC permits UWs consisting of bare English headwords without any constraints that would disambiguate their meaning. The U++ UWs always include some ontological constraints and list arguments of predicates.

The second UNL resource — UNLArium UNLdic (http://www.unlweb.net/unlarium/) is based on the Wordnet and we would criticize it for using numerical synset codes form Wordnet directly as UWs to identify concepts. The numbers merely refer to the size of the offset in bytes from the beginning of a Wordnet data file to the start of the synset record. They change between versions of Princeton Wordnet. Sticking to the numbers would severely limit the range of possible concepts in the resource, so UNLarium supplements them with proper UWs for concepts not directly linkable to the Wordnet. UNLdic is license compatible and it is relatively easy to exchange data between our projects.

4.2. Lexical networks

UNLDC contains a lot of data imported from Princeton Wordnet and reuses its synonymy and antonymy links. Even parts that were left out may be imported later. However even the English section of UNLDC is not an exact copy of the Wordnet and includes some new data, in particular the description of prepositions and conjunctions. More differences between UNLDC and Wordnet were described in [Dikonov V., Boguslavsky I., 2009].

Local dictionaries of other languages, e.g. French or Spanish, can be created from open Wordnets of those languages. The data contained in UNLDC and other linked resources together makes up a superset of a typical Wordnet. It is possible to generate new Wordnets from UNL data. For example, there is still no open and free Russian Wordnet that would be larger than 30,000 words. The already existing semantic and ontology relations make it possible to join registered Russian words into synsets and form a new open Russian Wordnet. The French local dictionary might fill in some gaps in WOLF, etc.

There are other projects of multilingual lexical networks consisting of words and translation links between them. Such projects usually lack semantic annotation, but their data can be used to extend the coverage of the semantic dictionary. One example is the PanLex project [Baldwin et al, 2010]. There are several Internet sites offering croudsourced multilingual dictionaries, e. g. Wiktionary, freedict, etc. All of them collect potentially useful data that can be used in future.

The existence of other projects aiming to integrate lexical data available in the Internet, e. g. BabelNet proves a widespread interest and importance of resource integration for the development of applied linguistics.

References

1. *Baldwin T., Pool J., Colowick S.* PanLex and LEXTRACT: Translating all Words of all Languages of the World, 2010
2. *Boguslavsky I., Cardeñosa J., Gallardo C., Iraola L.* The UNL Initiative: An Overview. Computational Linguistics and Intelligent Text Processing, 2005
3. *Boguslavsky I., Dikonov V.* Universal Dictionary of Concepts [Universal'nyj slovar' konceptov] Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2009" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009"]. Bekasovo, 2009 M.: RGGU, 2009. Issue. 8(15). pp. 91–96. ISBN 978-5-7281-1102-3.
4. *Boguslavsky I.* Guidelines for UW construction, manuscript
5. *Dikonov V. G.* Modal Attributes in UNL [Atributy modal'nosti v UNL]. Sbornik trudov 32-oj Konferencii molodyh uchenyh i specialistov IPPI RAN "Informacionnye tehnologii i sistemy (ITiS'09)" [Proceedings of the 32-nd Conference "Information technologies and systems (ITiS'09)"], Bekasovo, 2009. pp. 230–237. ISBN 978-5-901158-11-1.

6. *Dikonov V.* Semantic Network of the UNL Dictionary of Concepts. Proceedings of the SENSE Workshop on conceptual Structures for Extracting Natural language Semantics
7. *Iraola L.* Using WordNet for linking UWs to the UNL UW. International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies, Alexandria, EGYPT, 2003
8. *Nguyen Hong-Thai, Boitet C., Sérasset G.* PIVAX, an online contributive lexical data base for heterogeneous MT systems using a lexical pivot, SNLP-2007, Bangkok, 2007
9. *Pease, A.* (2011). Ontology: A Practical Guide. Articulate Software Press, Angwin, CA. ISBN 978-1-889455-10-5.
10. *Sagot B., Fišer D.* Building a free French wordnet from multilingual resources. Ontolex 2008, Marrakech, Maroc, 2008
11. *UNL* Specification 2005, available at: http://www.undl.org/unlsys/unl/unl2005/