Bulgarian and English Semantic Dictionaries for the Purposes of Information Retrieval

Max Silberztein – Universite de Franche-Comte
Svetla Koeva – Bulgarian Academy of Sciences

Abstract

The paper presents broad conception for stemming - as a mutual correspondence between word-form paradigms of all literals belonging to the synonymous sets constituting a given WordNet relation. The implementation includes the association of literals from the Bulgarian and English WordNets with the corresponding super-lemmas and inflection types in Semantic Dictionaries. The Semantic Dictionaries have been constructed with the NooJ linguistic development environment. NooJ dictionaries contain indistinctly simple or compound words thanks to an inflection system that can process both simple and compound words' inflectional morphology in a unified way. Moreover, NooJ can provide linking of all word forms associated with an equivalent super-lemma. When the super-lemma corresponds to a given semantic relation between words, a semantic stemming can be accomplished.

1 Introduction

The goals of the presented investigation[1] are directed to the implementation of natural language semantic relations in Information retrieval systems. This involves broad conception for stemming - as a mutual correspondence between word-form paradigms of all literals belonging to the synonymous sets constituting a given WordNet relation. The "semantic" stemming requires working out of the following tasks:

- to provide compete formalization of the inflection of simple and compound literals included in the Bulgarian and English WordNet structures;
- to create specialized Semantic Dictionaries for Bulgarian and English based on WordNet semantic relations.

Both tasks have been implemented with the NooJ linguistic development environment [7].

[1]The reported work is part of the Joint research RILA project *Information retrieval based on semantic relations* between LASELDI, Universit de Franche-Comt, and Department of Computational Linguistics, IBL, Bulgarian Academy of Sciences.

2 NooJ dictionaries

NooJ dictionaries contain indistinctly simple or compound words thanks to an inflection system that can process both simple and compound words' inflectional morphology in a unified way.

2.1 Merging simple and compound words

For instance, the two following lexical entries:
academic program,N+FLEX=APPLE
window,N+FLEX=APPLE
inflect the same way (they take an 's' in the plural). Therefore, they are both associated with the same inflectional class: APPLE. The class APPLE is defined by the following expression:
APPLE = $\langle E \rangle$/singular + s/plural;
that states that if one adds nothing to the lexical entry ($\langle E \rangle$ is the empty string), one gets the singular form ("singular"); if one adds an "s" to the end of the lexical entry, one gets the plural form ("plural"). NooJ's inflectional engine is equivalent to a stack automaton. It uses a dozen default commands that operate on the suffix of each lexical entry:

- $\langle B \rangle$: equivalent to the keyboard key "Backspace"
- $\langle D \rangle$: Duplicate current character
- $\langle E \rangle$: Empty string
- $\langle L \rangle$: equivalent to the keyboard key "Left arrow"
- $\langle N \rangle$: go to end of Next word form
- $\langle P \rangle$: go to end of Previous word form
- $\langle R \rangle$: equivalent to the keyboard key "Right arrow"
- $\langle S \rangle$: equivalent to the keyboard "delete" key

Users can override these commands, and add their own.
NooJ is capable of inflecting compounds. For instance, the class "ACTOFGOD" is defined by the following expression:
ACTOFGOD = $\langle E \rangle$/singular + $\langle P \rangle \langle W \rangle$s/plural;
The operator $\langle PW \rangle$ stands for: "go to the end of the first component of the lexical entry". Note that the following three entries are associated with this class, even though their length is different:
bag of tricks,N+FLX=ACTOFGOD
balance of payment deficit,N+FLX=ACTOFGOD
member of the opposite sex,N+FLX=ACTOFGOD
In the same manner, even though the lexical entries: *blank piece of paper, last line of defence, family history of cancer, sexual harassment in the work place,*

etc. have different lengths, they can be associated with a unique inflectional class because the inflection is carried by the same component (the second one). Agreements between components of a compound can be described as well. For instance, the following inflectional expression formalizes the agreement between the two components of compounds such as journeyman carpenter:
$\langle E \rangle$/singular + s$\langle P \rangle\langle B \rangle$2en/plural
To get the plural form, add an "s" to the end of the compound, then go back to the previous ($\langle P \rangle$) component, delete the two last characters ($\langle B \rangle$2), and add the suffix "en". In conclusion, NooJ can process simple and compound words' inflection completely automatically. This has allowed us to unify the description of simple and compound words in WordNet type dictionaries.

2.2 Dictionary and Inflection

NooJ dictionaries are directly compiled into a Finite-State Transducer, in which all the relevant inflectional paradigms are stored as well. This characteristic is very important for our project: it gives NooJ the ability to perform morphological operations during parse time. NooJ can link any inflected form, to any other inflected form represented in its transducer. Thus, NooJ can perform complex transformations within texts. For instance, it is now possible to replace a certain conjugated verb with its past participle form, and vice-versa, within a particular text:
John eats the apple $\leftrightarrow$ the apple is eaten by John
Using this new functionality will enhance several current NLP applications, such as automatic translation and information retrieval applications.

2.3 Property definition and types

NooJ dictionaries can be displayed either in list ("free") form, or in table ("typed") form. This requires that features of the dictionary be typed, so that all the values of a common property can be regrouped in one column. We also need to state the number of relevant properties for each category of word (e.g. Tense for Verbs, Number for Nouns, etc.), in order to distinguish absent default values, from irrelevant ones. This is done via a "Property Definition" file that contains rules such as:
NDistribution = Hum + Conc + Abst ;
NGender = m + f ;
NNumber = s + p ;
...
VTense = Present + Futur +... ;
VPers = 1 + 2 + 3 ;
VNumber = s + p ;
...
The next figure displays a NooJ dictionary in table form:

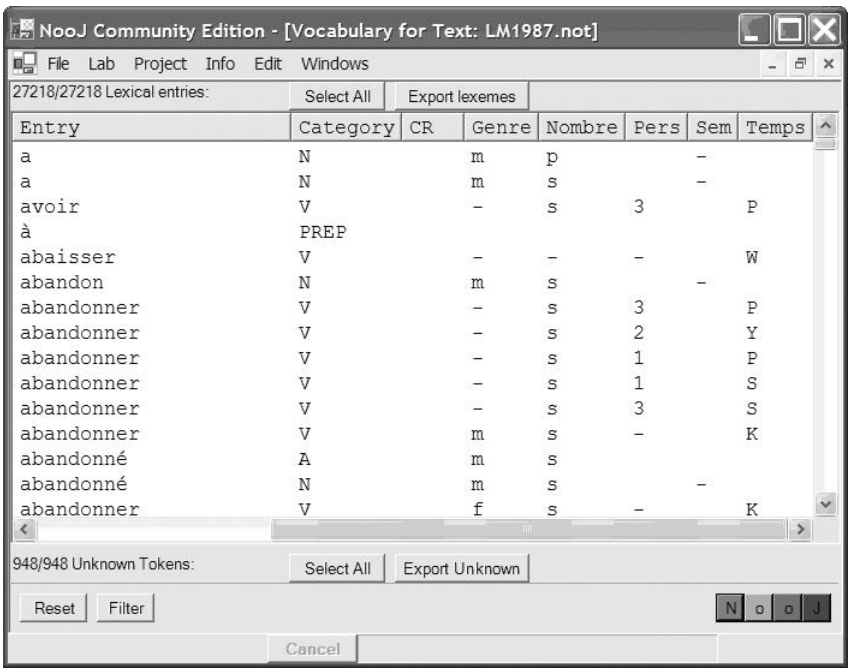

Figure 1 Table view for a NooJ dictionary

Note that all features in a NooJ dictionary do not have to be typed; if NooJ does not know the type of a feature, it will simply display it as a column header, and enter the "+" (if the feature is present) and "-" (if absent) values accordingly. A given property may be associated with more than one category (e.g. Number is relevant both for nouns and verbs). But NooJ checks that one feature (e.g. "+p") does not correspond to more than one property (e.g. "Gender" and "Tense"). NooJ distinguishes default properties from irrelevant ones. Moreover, associating NooJ lexical features with typed properties opens up possibilities for implementing unification mechanisms in the future.

2.4 Bulgarian Grammatical Dictionary

The grammatical information included in the Bulgarian Grammatical Dictionary (BGD) is divided into three types [2]: category information that describes lemmas and indicates the words clustering into grammatical classes (Noun, Verb, Adjective, Pronoun, Numeral, and Other); paradigmatic information that also characterizes lemmas and shows the grouping of words into grammatical subclasses, i. e. - Personal, Transitive, Perfective for verbs, Common, Proper for nouns, etc.; and grammatical information that determines the formation of word forms and shows the classification of words into grammatical types according to their inflection, conjugation, sound and accent alternations, etc. The BGD is a list of lemmas where each entry is associated with a label [4]. The label itself represents the grammatical class and subclass to which the respective lemma

belongs and contains a unique number that shows the grammatical type. All words in the language that belong to the same grammatical class, subclass and have an identical set of endings and sound / stress alternations are associated with one and the same label. Each label is connected with the corresponding formal description of endings and alternations.
The inflectional engine used is equivalent to a stack automaton. Despite the existence of some differences in the format, the BGD represents a kind of DELAS dictionary, and it is compiled into a Finite-State Transducer. The BGD which already contains over 85000 lemmas has a parallel version in NooJ format where dictionary labels are transliterated and formal descriptions are transformed into the NooJ formal apparatus.

3 WordNet

The global WordNet [1]; [6] is an extensive network of synonymous sets and the semantic relations existing between them, enabling cross-language references between equivalent sets of words in different languages [9]. The Bulgarian wordnet (BulNet) has been initially developed in the framework of the project *BalkaNet – a multilingual semantic network for the Balkan Languages* which has been aimed at the creation of a semantic and lexical network of the Balkan languages [8].

	Bg N	Bg V	Bg Adj	Bg Adv	Bg Total
Synsets	15 508	4 421	4 027	449	24 405
Literals	27 772	15 701	7 017	1 094	51 584
Graphic-words	22 381	8 860	5 040	817	37 098
ILR	25 577	11 143	6 931	815	44 466

Table 1: The distribution of Bulgarian synsets into parts of speech

The Bulgarian WordNet [3] models nouns, verbs, adjectives, and (occasionally)

	En N	En V	En Adj	En Adv	En Total
Synsets	79 689	13 508	18563	3 664	115 424
Literals	141 691	24 632	31 016	5 808	203 147
Graphic-words	114 649	11 306	21 437	4 660	152 052
ILR	129 983	36 457	34 880	3 628	204 948

Table 2: The distribution of synsets in English Wordnet 2.0

adverbs, and contains already 24405 word senses (towards 1.09.2005), where 51584 literals have been included (the ratio is 2,11). The distribution of Bulgarian and English synsets across different parts of speech is shown in Tables 1 and 2.

3.1 Wordnet structure

Every synset encodes the equivalence relation between several literals (at least one has to be present), having a unique meaning (specified in the SENSE tag value), belonging to one and the same part of speech (specified in the POS tag value), and expressing the same lexical meaning (defined in the DEF tag value). Each synset is related to the corresponding synset in the English Wordet2.0 via its identification number ID. There has to be at least one language-internal relation (there could be more) between a synset and another synset in the monolingual data base. There could also be several optional tags encoding usage, some stylistic, morphological or syntactical features, etc.

3.2 Inflecting Wordnet

Literals included in the wordnet structure can be either simple words or compounds i.e. the English synset *car:2, railcar:1, railway car:1, railroad car:1* with the definition "it a wheeled vehicle adapted to the rails of railroad" corresponds to Bulgarian synset *vagon:1*; the Bulgarian synset *hol:1 salon:1 balna zala:1* with the definition "the large room of a manor or castle" corresponds to the English one *manor hall:1, hall:5*. Comparing to lemmas compounds have their own inflective rules. In order to merge the language data existing in BulNet and BGD it was decided to assign an additional grammatical note to each literal thus linking it with the BGD lemma's label [5]. All labels for BGD entry forms that are found in the BulNet have been entered as values of the LNOTE (lexical note) grammatical tag in the XML format. Most of the literals which were not recognized are either specialized terms that have no place in a grammatical dictionary of the common lexis (often written in Latin) or compounds. The contradictory cases where two or more labels were associated with one and the same literal are solved manually. The classification of compounds according to different inflectional types is under development.

3.3 WordNet relations

The major part of the relations encoded in the Bulgarian WordNet is semantic relations. There are also some morpho-semantic relations, some morphological (derivational) relations, and some extralinguistic ones. WordNet relations of equivalence, inheritance, similarity, and thematic domains affiliations are of interest to the Information retrieval purposes. Those are: synonymy; hypernymy, meronymy, similar to, verb group, also see, and category domain.

Synonymy

Synonymy is a semantic relation of equivalence (reflexive, symmetric, and transitive) between literals belonging to one and the same part of speech. In Princeton WordNet the substitution criteria for synonymy is mainly adopted: "two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value" [6]. Thus the relation implies that one synonym may substitute another (synonym) in a context and vice versa.

The consequences from such an approach are at least two - not only the exact synonymy is included in the data base (a context is not every context). Second, it is easy to find contexts in which words are interchangeable, but still denoting different concepts (for example hypernyms and hyponyms), and there are many words which have similar meanings and by definition they are synonyms but are hardly interchangeable in any context due to different reasons - syntactic, stylistic, etc. (for example an obsolete and a common word).

Hypernymy

Hypernymy (Hyponymy) is an inverse, asymmetric, and transitive relation between synsets, which correspond to the notion of class-inclusion between synsets belonging to one and the same part of speech. The relation implies that the hypernym may substitute for the hyponym in a context but not the other way round. Hypernymy is a transitive relation: e.g. being a kind of *cvete (flower), roza (rose)* has inherited not only all semantic features of *cvete (flower)*, but also those of its superordinates: *rastenie (plant), zhiv organizam (organism)*, etc.

Meronymy

Meronymy (Holonymy) is inverse, asymmetric, and transitive relation which link synsets denoting wholes with those denoting their parts. The Part meronymy typically relates components to their wholes. In BulNet we restrict the Part relation to the components that are topologically included one in the other with physical attachment: *book* is a part of *library*, *library* is a part of *building*, **book* is not a part of *building*, only *library - building* relation is encoded as Part meronymy. The Member meronymy is a relation between sets and their members i.e. *football player - football team - football league*, *football player* is a member of a *football team*, *football team* is a member of *football league*, as well as a *football player* is a member of *football league*. The Portion meronymy is between wholes and their portions i.e. *crumb of bread - slice of bread - loaf of bread*; *crumb of bread* is a portion of a *slice of bread*, *slice of bread* is a portion of *loaf of bread*, as well as *crumb of bread* is a portion of *loaf of bread*.

Relations of equivalence

Similar to is a symmetric relation between similar adjectival synsets. i.e. the synset *nice:1* with a gloss "pleasant or pleasing or agreeable in nature or appearance" is in a Similar to relation with the synnset *good:7* defined "as agreeable or pleasing". Verb group is a symmetric relation between semantically related verb synsets: the synset *wash:9, wash out:4, wash off:1, wash away:2* with meaning "remove by the application of water or other liquid and soap or some other cleaning agent" is in a Verb group relation with the synset *wash:1, rinse:2* with a definition "clean with some chemical process". Also see is a symmetric relation between synsets - verbs or adjectives, that are close in meaning i.e. *beautiful:1* defined as "delighting the senses or exciting intellectual or emotional admiration" is in a relation Also see with *attractive:1* defined as "pleasing to the eye or mind especially through beauty or charm".

Category domain

Category domain is an asymmetric extralinguistic relation between synsets denoting a concept and the sphere of knowledge it belongs to i.e. the synset *alibi:1*

with the definition "a defence by an accused person purporting to show that he or she could not have committed the crime in question" is in a Category domain relation with the synset *law:2, jurisprudence:2* defined as "the collection of rules imposed by authority".

4 Semantic stemming

4.1 Multi-fields dictionaries

The number of fields of NooJ dictionaries is no limited to one. NooJ dictionaries can contain entries associated with a "super-lemma", that can be an orthographical variant, the translation in another language, a synonymous entry or an hyperonym. For instance, consider the following lexical entries:
U.N.,United Nations,N+Org
czar,tsar,N+FLX=Pen
The first entry (U.N.) is associated with super-lemma "United Nations"; it does not inflect. This entry is similar to a DELACF entry. The second entry (czar) is associated with super-lemma "tsar"; it inflects according to the paradigm "Pen" (i.e. takes an 's' in the plural). Being able to associate words with super-lemmas, i.e. words that do not necessarily correspond to their inflectional lemma ("czar" is czars's lemma, not "tsar") opens up a new range of applications.

4.2 Semantic Dictionaries

The Semantic Dictionaries are designed using the WordNet structures (enumerated relations), on the one hand, and the respective inflectional dictionaries, on the other hand. There are two types of relations in the wordnet - symmetric (as synonymy) and asymmetric (as hypernymy) which determine the two approaches with super-lemma association. For the symmetric relations the super-lemma in Semantic Dictionaries is considered as the IDentification number of a given synonymous set.
author,ENG20-10090311-n,N+FLX=APPLE
writer,ENG20-10090311-n,N+FLX=APPLE
For the asymmetric relations the formalization is in the direction from the more concrete to more global concept (thus the super-lemma is the ID of the highest synonymous set in the hierarchy), but the other way is also possible. The main applications of the Semantic Dictionaries are directed towards Information retrieval by means of: semantic equivalence with synonymy dictionaries, semantic specification with hyperonymy and meronymy dictionaries, Information retrieval by means of similarity and thematic domains affiliations. The Semantic Dictionaries provide retrieve of all word-forms of all literals belonging to the synonymous sets constituting a given WordNet relation (Figure 2).

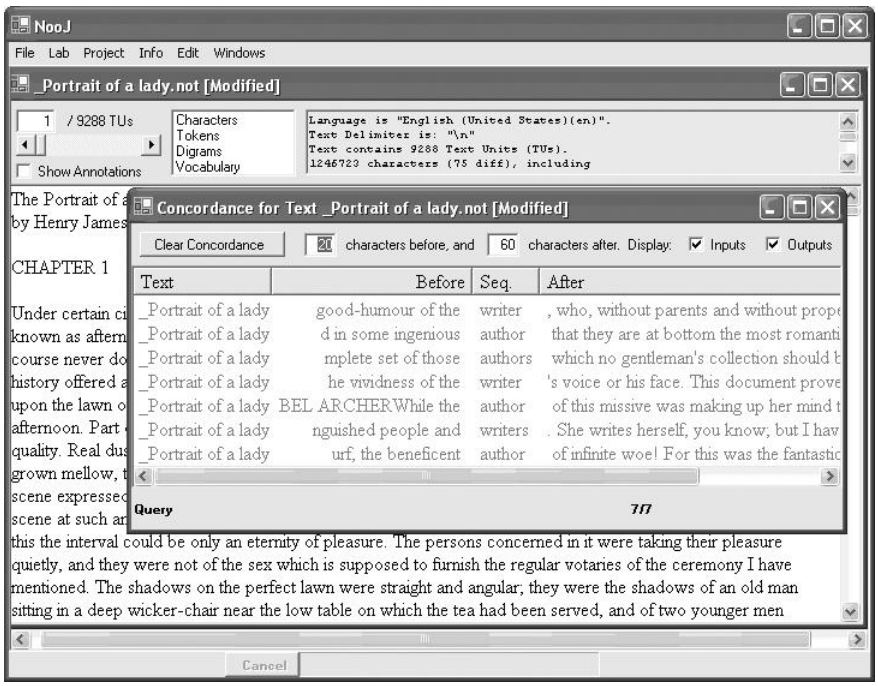

Figure 2 Semantic stemming

5 Conclusions and future directions

The Multi-fields dictionaries can provide Information retrieval by means of semantic equivalence with synonymy dictionaries, by means of semantic specification with hyperonymy and meronymy dictionaries, by means of similarity relations, and by means of thematic domains affiliations. Future work is directed to the extensions and enhancements of the Semantic Dictionaries:

- Extension of the dictionaries coverage;
- Addition of other semantic relations;
- Inclusion of additional information to the entries.
- Integration of multilingual semantic extraction with NooJ using the Inter-Lingual-Index relation.

References

[1] Fellbaum, C. (ed.). WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press, 1998.

[2] Koeva S., Bulgarian Grammatical dictionary. Organization of the language data, Bulgarian language, 1998, vol. 6: 49-58.

[3] Koeva S., T. Tinchev and S. Mihov Bulgarian Wordnet-Structure and Validation in: Romanian Journal of Information Science and Technology, Volume 7, No. 1-2, 2004: 61-78.

[4] S. Koeva Modern language technologies – applications and perspectives, in: Lows of/for language, Hejzal, Sofia, 2004, 111- 157

[5] S. Koeva Validating Bulgarian WordNet using grammatical information in: Proceedings from Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, Gteborg University, 2004, 80-82.

[6] Miller G. A. Introduction to WordNet: An On-Line Lexical Database. In "International Journal of Lexicography", Miller G.A., Beckwidth R., Fellbaum C., Gross D., Miller K.J. Vol. 3, No. 4, 1990, 235–244.

[7] Silberztein M. NooJ: an oriented object approach. In INTEX pour la Linguistique et le traitement automatique des langues. Les Cahiers de la MSH Ledoux. Presses Universitaires de Franche-Comte: Besancon, 2004.

[8] Stamou S., K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufis, 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.

[9] Vossen P. (ed.) EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer Academic Publishers, Dordrecht. 1999.