Proceedings of the 2007 International NooJ Conference

Proceedings of the 2007 International NooJ Conference

Edited by

Xavier Blanco and Max Silberztein

Cambridge Scholars Publishing

Proceedings of the 2007 International NooJ Conference, Edited by Xavier Blanco and Max Silberztein

This book first published 2008

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-0053-8, ISBN (13): 978-1-4438-0053-2

TABLE OF CONTENTS

INTRODUCTION

XAVIER BLANCO AND MAX SILBERZTEIN

This volume contains a selection of 18 papers, chosen from among the 38 papers that were presented at the *2007 NooJ conference*, Autonomous University of Barcelona, June 7-9, 2007.

NooJ is a linguistic development environment that allows linguists to formalize a wide gamut of linguistic phenomena, and then test, adapt, share and accumulate each elementary description to build linguistic “modules”, i.e. structured libraries of linguistic resources. NooJ is also used as a corpus processor that can launch sophisticated queries over large corpora in order to produce various results (concordances, statistical analyses, information extraction, etc). Finally, NooJ’s linguistic engine is integrated in several research centers and software companies’ software in order to build various Natural Language Processing applications.

NooJ, released in 2002, is Max Silberztein’s new linguistic-friendly formalization platform([1]). As opposed to most other linguistic engineering software which specialize in the formalization of one or two levels of linguistic phenomena, NooJ offers a unified platform for the development of any linguistic engineering project. Since 1993, a large community of users who share the same scientific and technical goals, follow the same methodology and share software tools and linguistic resources has grown up around NooJ.

NooJ proposes a methodology and provides specific tools for the formalization of five levels of linguistic phenomena: orthography, morphology, lexicon, syntax and semantics. The keyword here is *specific*: as opposed to other linguistic models and platforms, NooJ provides tools adapted specifically to each phenomenon to be formalized, so that the description of each linguistic phenomenon is as natural (i.e. easily understood by linguists), and as thorough as possible. For each level of

[1] The first attempt, INTEX, was released in 1993. INTEX was basically an integrated version of a software toolbox used to perform finite-state processing on dictionaries and texts; this toolbox was constructed by Max Silberztein for his 1989 PhD thesis work, see http://intex.univ-fcomte.fr.

formalization, NooJ proposes a corresponding parser that can be used to test each piece of the linguistic formalization over large corpora. Hence, NooJ includes an orthographic parser (basically a sophisticated, highly flexible tokenizer), a morphological parser that can perform inflectional and derivational analyses, a set of tools to look up complex lexicons, a syntactic parser that produces structured annotations and a semantic parser that performs automatic transformations. All parsers communicate thanks to a Text Annotation Structure (TAS) that stores results produced by each parser, as well as all unsolved ambiguities. Max Silberztein's paper: "Complex Annotations with NooJ", describes the TAS, its function, and how it is used.

Thus, building a "NooJ module" involves constructing five types of linguistic resources:

– The orthographical level is where one deals with the absence of word delimiters in Asian languages, the absence of vowels in texts written in Semitic languages, intonation marks inside words in Armenian, tokenization of words in agglutinated languages, the properties of accented letters and diacritics specific to each language, the multiple functions of the period, dash and apostrophe, the idiosyncratic rules used for each alphabetic order, etc.

Huei-Chi Lin's paper "Treatment of Chinese Orthographical and Lexical Variants with NooJ" describes how the massive character variation in Chinese texts can be formalized at the orthographical, morphological and lexical levels.

– The morphological level contains inflectional morphology (e.g. how to conjugate verbs), derivational morphology (e.g. how to construct nouns from verbs), and productive morphology (e.g. how to construct new terms).

Andrei Filip's paper "Building NooJ Inflection Graphs for the Morphological Description of Nouns in Romanian" and Odile Piton and Klara Lagji's paper "Morphological Study of Albanian words" show morphological phenomena that occur specifically in these languages, and how the authors formalized them.

– The lexical level provides an environment to develop lexicons that describe all elements of a language's vocabulary in a unified way, including simple words, multi-word units and discontinuous expressions. Each lexical entry might be linked with resources developed at the other levels of description (orthography, morphology, syntactic and semantic levels), so that, for instance, a word will automatically be linked to its spelling and morphological variant forms, and it will be associated with information relevant to the syntactic and semantic levels.

Anabela Barreiro's paper "Port4NooJ: Portuguese Linguistic Module and Bilingual Resources for Machine Translation", Krzysztof Bogacki's paper "Polish Module for NooJ" and Zoe Gavriilidou et alii's paper "The New Greek Module: Morphosemantic Issues" describe the construction of a dictionary for NooJ and its corresponding linguistic resources (orthographical and morphological).

– The syntactic level is where one describes local syntactic phenomena such as the order of preverbal particles in French (e.g. *je la lui donne*), the multiple ways to express a date (e.g. *Monday, April 13th 2008*), complex sequences of determiners (e.g. *most groups of my friends*), various types of morphological agreements, etc. Users can also build, test and debug Context-Free Grammars (CFG) in the form of "libraries", as well as Augmented Transition Networks (ATN) in which grammatical rules and complex agreement constraints are combined to reduce structural ambiguities.

Frozen and semi-frozen expressions are some of the most complex linguistic phenomena to formalize because they are numerous (an ad hoc solution is not sufficient) and their behavior needs to be formalized both at the lexical and at the syntactic levels. Peter Machonis's paper "NooJ: a Practical Method for Parsing Phrasal Verbs" shows how he implemented a lexicon-grammar of English phrasal verbs as a pair (lexicon, grammar). Maria Todorova's paper "Morpho-Syntactic Properties of Bulgarian Verbal Idiomatic Expressions", on the other hand, focuses on semi-frozen expressions and shows how to implement them with local syntactic grammars.

– The semantic level is where linguists can build semantic extractors (e.g. Named Entities) and formalize expressions and sentences alternations. In NooJ, alternations are seen as *transformations* that can be applied to texts automatically. A typical expression transformation is to produce the verbal paraphrase of a noun phrase (e.g. *John's gift to Mary → John gave something to Mary*); a typical sentence transformation is to produce the passive form of a sentence (*John ate the apple → The apple was eaten by John*), etc. NooJ's transformational capabilities are also used to produce semantic analyses of expressions, in the form of logical formula such as "EAT (John, Apple)". NooJ's transformational capabilities are also used to perform local translations of expressions or sentences (e.g. *John a mangé la pomme*).

Sandra Gucul-Milojević et alii's paper "Usage of NooJ Graphs and Annotation for Information Extraction" presents an elegant library of syntactic local grammars used to automatically extract person names and temporal expressions from a large Serbian corpus of journalistic texts.

These two grammars are then combined into a complex query that extracts automatically from texts expressions in which a person says something at a certain date. Ivelina Stoyanova et alii's paper "The Treatment of Named Entities in Machine Translation" presents a particularly exciting new application that adds translation capabilities to a named entity extractor implemented by a library of local grammars.

Svetla Koeva's paper "Syntactic Alternations" presents a methodology to formalize alternations, which yields to classifying them in three classes: alternations that alter some lexical properties of the verb, alternations that involve only pure syntactic transformations, and alternations that imply some semantic analysis in order to process implicit arguments.

NooJ is also widely used as a sophisticated corpus processor, i.e. to parse large corpora with complex queries.

Alexandrov et alii's contribution "NooJ Applications for Document Clustering and Corpus Linguistics" deals with several applications of NooJ in corpus mining, including autorship attribution. Natalia Ponomareva et alii's paper "Regression Model for Politeness Estimation Trained on Examples" describes a system capable of analyzing dialogue interactions automatically. The paper focus on "politeness" indicators, which are words (e.g. *por favor*) or expressions (e.g. *me podría*) that are counted in order to estimate the level of politeness of the interaction. Magali Bigey's paper "Specific Electronic Dictionaries and Literary Corpora" shows how to perform a complex analysis of a large literary corpus by combining semantically two homogeneous dictionaries: a dictionary of body part nouns (e.g. *head, lips*) and a class of feeling verbs (e.g. *to deceive, to enchant*).

Finally: NooJ's linguistic engine has been integrated into a number of Natural Language Processing software applications. Slim Mesfar's paper "NooJ4Web: an On-line Concordance Service" presents a new service available on-line, that allows linguists to launch sophisticated queries over large corpora of texts in Arabic, English and French. This concordancer uses NooJ's linguistic engine. Ranka Stanković et alii's paper "The NooJ System as Module within an Integrated Language Processing Environment" presents a new linguistic platform, WS4LR, also based on NooJ's linguistic engine, capable of applying multilingual Wordnet-type dictionaries in order to launch queries on parallel multilingual texts.

Although Johannes Stiehler does not use NooJ, his invited paper "Dynamic Fact Extraction through Structural Search" presents a software architecture based on annotations which is very similar to NooJ's TAS.

We think that the reader will apreciate the importance of this volume, both for the value of each linguistic formalization, the underlying research framework, the pedagogical value of the experiments presented here, as well as the construction of new Natural Language Processing applications.

NooJ Applications for Document Clustering and Corpus Linguistics

Mikhail Alexandrov, Xavier Blanco, Olga Mitrofanova and Victor Zakharov

Abstract. NooJ is a well-known tool for professionals giving them an opportunity to reveal linguistics patterns in textual documents. This paper considers NooJ as a necessary element of preprocessing in the tasks dealing with data organization and corpus mining. In the first case NooJ provides keyword-based and style-based indexing, and in the second case graphemic analysis, morphological tagging, grammatical disambiguation, etc. We demonstrate various examples, where NooJ proves to be effective: in sentiment analysis, problems of authorship, clustering words/documents, word sense disambiguation and computer modeling of morphology.

Key words: NooJ, indexing, clustering, corpus linguistics

Introduction

NooJ is a well-known language dependent software, which uses lexical resources and morphological dictionaries of a given language for detailed linguistic research. Such reseach may include lexicographic and morphological analysis, classification of style and genre and other tasks being of great interest for professional linguists and philologists. Alongside with traditional linguistic research NooJ can be effectively used in the tasks of information retrieval and corpus analysis, where NooJ is an important element of preprocessing. Such preprocessing can include delicate style-based text parameterization, statistics of hidden relations between various lexicographic entries, etc. In the paper we give some examples of such applications.

The paper is organized as follows. Section 2 is devoted to NooJ applications in clustering. Section 3 describes NooJ applications in corpus linguistics. Conclusions are drawn in Section 4.

1. Clustering Procedures and NooJ

1.1 Clustering and Indexing

Clustering is one of two principal approaches for grouping data sets, the other approach being classification. Table 1 contains short description of these two types of grouping data [Alexandrov, Makagonov, 2007].

Generally the following three objectives of clustering are considered as the principal ones: structuring object set, searching interesting objects, knowledge discovery. As for knowledge discovery, it should be noted that clustering is the main approach in this area. It is really so: when objects or attributes are joined in some groups, then an expert can think about the reasons of such grouping. What circumstances obliged data to be joined together?

Table 1 Clustering and Classification

Type of grouping data	Synonyms	Description
Clustering	- Classification without teacher - Unsupervised classification	Absence of patterns or descriptions of classes. So, the results are determined by internal nature of data . Obviously, the procedure of clustering is possible if the number of objects is more than one. Classes (groups) are named *clusters*
Classification	- Classification with teacher - Supervised classification	Presence of patterns or descriptions of classes. So, the results are determined by external reasons. Obviously, classification is possible even for *one object.* Classes (groups) are named *categories*

It is evident that all procedures related with grouping need some quantitative measure of similarity/distance between objects. It means that all objects must be parameterized (described) by means of their attributes. Speaking about attributes we mean here both lexicographic and stylistic text markers.

Usually types and number of attributes involved in text classification are defined by the given categories. But in clustering we often have contradictory situation as attributes are not fixed. In this case we ourselves should select attributes which could reveal essential relations between objects. NooJ proves to be very useful for such selection. Just for this reason NooJ is more suited for clustering than for classification.

Attribute selection in text processing is named indexing. At present many methods of keyword-based indexing are developed and used in practice [Baeza-Yates, 1999]. NooJ gives the possibility to realize style-based indexing.

1.2 Methods of Clustering

For a moment, there are dozens of methods and their modifications in cluster analysis, which can satisfy practically all necessities of users. The most popular ones are *K*-means, oriented on the structures of spherical form, and Nearest Neighbor (NN), oriented on the extended structures of chain form [Hartigan, 1975]. In recent years MajorClust method proved to be very successful in applications related with text information retrival [Stein, Mayer zu Eissen, 2002, 2003]. MajorClust belongs to the group of density-based methods and is not related with any certain data structure. This method has the following advantages over the mentioned two methods:

– MajorClust distributes objects to clusters in such a way that the similarity of an object to the assigned cluster exceeds its similarity to any other cluster. This natural criterion provides the grouping of objects, which better corresponds to the users' intuitive representation. Neither *K*-means nor NN methods possess such optimization property: they do not evaluate similarity of clusters.

– MajorClust determines the number of clusters automatically and in all cases tends to reduce this number. *K*-means requires the number of cluster to be given, and NN does not determine this number at all: cutting of the dendrite is performed by the user.

It is not a joke but the number of clustering methods now is a little bit more than the number of researchers working in this area. So, the problem does not consist in searching the best method for all cases. The problem consists in searching the method being relevant for the data under consideration. Only user knows what methods are the best for his/her own data. What should we do first of all? We need to pay attention to the choice of indexes (parameters) and measure of similarity/distance. They should be adequate with regard to a given problem and a given data set.

Frequently the results are bad because of the bad indexes and bad measure but not the bad method!

1.3 Style-based clustering for dialogs processing

Cluster analysis of dialogs with transport directory service allows to reveal typical scenarios of dialogs, which are useful for development of automatic dialog systems. The typical dialog is presented in the Table 2. Here DI means directory inquires and US means a user.

For clustering dialogs we introduced various indicators reflecting both the charateristics of trip and the passanger's behavior such as point of destination, type of train, urgency of trip, level of politeness, length of talk, etc. Clustering was completed by MajorClust and its results are presented here [Alexandrov, Sanchis, Rosso]:

- Cluster 1 (31 objects). No urgent trips, no night trips (only 20%), only ordinary politeness.
- Cluster 2 (44 objects). Urgent trips or defined days of trips (95%), advanced level of politeness (85%).
- Cluster 3 (12 objects). Only small and middle cities, no urgent trips, one-way trips (75%), short talking (85%), the highest level of politeness.
- Cluster 4 (13 objects). Only small and local cities, undefined days of trip, one-way trips (75%), no night trips (only 15%), short talking (75%), advanced level of politeness.

It should be noted that the contents of clusters 3 and 4 were expected: dwellers of small cities are usually more polite that the dwellers of middle and large cities.

In this experiment we used NooJ for revealing indicators of politeness. NooJ detected the following 3 indicators: *first greeting, polite words* and *polite expressions* based on subjunctive mood. Just these indicators were used in subjectivity analysis described in the Section 3.5 [Ponomareva, Blanco, 2008].

Table 2 Example of real dialog between passengers and directory inquires

DI: Renfe, good day
US: Yes, good day
DI: Yes, well.
US: OK, could you inform me about the trains that go from here, from Barcelona to Valladolid?
DI: What day it will be?
US: The next Thursday.
DI: Let us to see. <PAUSE> on Thursday is off the one at thirteen, which come at twenty two hours to Valladolid
US: Are there any more?
DI: No, on Thursday only this one, eh?
US: Nothing more, say me, please.
DI: Exactly.
US: <CONTINUALLY> before the Wednesday or Thursday.
DI: The train will be exactly at evening, on Thursday or Friday it is off.
US: Thank you, bye

1.4 Style-based clustering in the problem of plagiarism/authorship

The open access to Internet resources proved to be the reason of numerous cases of plagiarism. Now such a topic is actively discussed in Internet community and recently the first Conference devoted to plagiarism analysis has been held [PAN 2007]. Obviously, the most popular indicators used in plagiarism analysis could be also used in the problem of authorship detection. 10 carefully selected style markers and about 10 part-of-speech features are suggested and elaborated in [Meyer zu Eissen, S., Stein, B., Kulig, M, 2007; Stein, B., Meyer zu Eissen,. S., 2007]. Other approaches to authorship detection are described in [Marusenko 1990]. All markers relevant for authorship detection are easy detected by NooJ.

Challenging research dealing with the well-known problem of authorship of Molière and Corneiller comedies was carried out [Labby, C., Labby, D., 2001], see also [Rodionova 2007, 2008]. The fact is the authorship of many Molière and Corneille dramatic works are definitely established. But there were a set of works whose authorship caused doubts. The Fig. 1 schematically presents this situation.

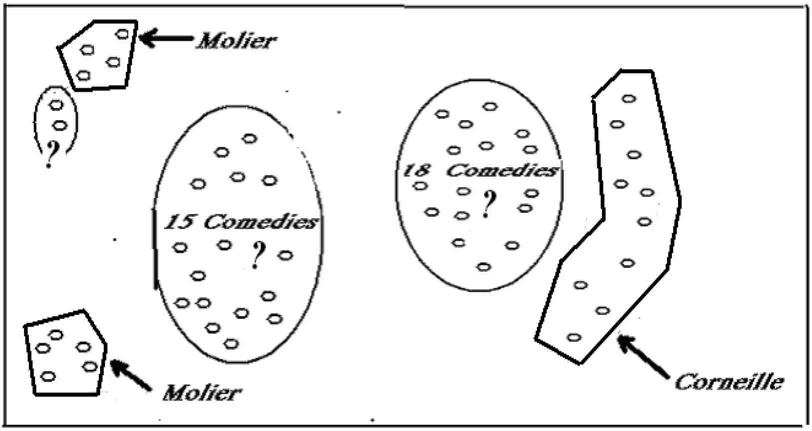

Fig. 1 Schematic position of Molière and Corneille comedies and also arguable works

According to this picture the following results can be formulated and justified:

-18 comedies of Molière should belong to Corneille

-15 comedies of Molière are weakly connected with all his other works. So, they can be written by two authors

-2 comedies of Corneille now are considered as works of Molière.

It should be said that during a certain period Molière and Corneille were close friends. In this research style-based indicators were manually estimated. NooJ could quicken this work to a considerable extent.

2. Corpus Processing and NooJ

2.1 Current State of Corpus Linguistics

Recently development of text corpora has been at the cutting edge of linguistic engineering. The core stage of corpus construction is the specialized procedure of text description, which implies introducing mark-up (tagging, annotation). In general, there are three levels of text description and, thus, three types of metadata, namely, quasi-bibliographic, structural and linguistic ones. The first level of text description includes introduction of a set of standard bibliographic data elements and a set of elements of literature and book science characterizing genre, style, the history of writing and publishing, etc. At the second level the following elements are added to the document structure: text, chapter, section, paragraph, sentence, phrase, word. The identification of some multiword units is fulfilled. Some special tags are produced for the beginning and for the end of a sentence and for the rest of the punctuation signs. Those data are relevant for morphological disambiguation, which is performed at the next level of text processing. The third level implies morphological tagging, i.e. lemmatization and assigning morphological characteristics to all tokens. Syntactic and semantic analysis as well as disambiguation may be carried out at this level.

Alongside with text description researches are to provide supplementary linguistic resources (grammars and dictionaries) aimed at effective multilevel analysis of texts and specialized tools (corpus managers) for extensive search in corpora. Morphosyntactic parsers as well as lexical databases are available for corpora developers. Corpus managers allow to perform such valuable procedures as creating subcorpora, context search (for lemmas and tokens), concordance construction, frequency lists construction, word bigrams search, measuring associative relations in collocations, etc. Irrespective of the operational scope, a significant parameter of any corpus manager is flexibility, i.e. compatibility with other linguistic tools, access to different types of texts and text data, parallel processing of a set of texts or subcorpora, etc. As for currently used corpus managers, most of them are corpus-dependent and are functionally restricted.

2.2 NooJ Applications in Russian Corpus Linguistics

In discussing linguistic resources and tools involved in processing of Russian texts for creating corpora of various types, the question arises– what tools can be used to process Russian texts at all stages of corpus construction? To meet the need, several software packages supplied with Russian linguistic modules have been developed: e.g.:

- AOT (www.aot.ru),
- MyStem (http://corpora.narod.ru/mystem/),
- StarLing (http://starling.rinet.ru/morpho.php?lan=en), etc.

Proper analysis of those tools shows that they perform all the required operations but their possibilities are not unlimited. From this point of view, NooJ seems to be a highly elaborated universal linguistic environment which allows to fulfill the whole set of corpora development tasks, namely, graphemic (premorphological) analysis, morphological tagging, grammatical disambiguation, syntactic and semantic processing of texts, quantitative analysis, etc. It is necessary to ascertain compatibility of existing Russian linguistic resources (grammars and dictionaries) and NooJ tools; to clear up technology of merging Russian-oriented tools and NooJ tools within a common research environment; to perform trial processing of Russian corpora with the help of NooJ.

2.3 Graphemic analysis and NooJ dictionaries

One of the vital problems to be solved in most NLP systems is graphemic analysis, which includes segmentation of the input text in terms of words, separators, etc. Determination and proper XML-tagging of various nonstandard and usually nonlexical elements of the input text is conducted simultaneously and involves formatting elements (e.g., strong text, italics, underlining, etc.); structural text elements (e.g., headers, indents, comments, etc.); various text nonverbal elements (e.g., numbers, dates in numerical formats, alpha-numerical complexes, etc.); names (e.g., personal names, patronymics, written as initials); foreign lexemes written in Latin alphabet, etc.

The major difficulty in graphemic analysis is processing of the hyphen that is of high frequency and which improper processing leads to a great number of errors.

A question arises if the inter-word hyphen treatment as a character is better than its treatment as a punctuation mark or vice versa. The inter-word hyphen treatment as a character facilitates the analysis of graphical words when their parts have a hyphen between them (e.g., *kto-to (someone), gde-nibud (somewhere), davnym-davno (long ago)*, etc.). However such a treatment leads to the loss of simplicity in the analysis of freely derived juxtaposed compounds (e.g., *starik-khudozhnik (aged artist), slovar-spravochnik (reference-book)*, etc.). Naturally, such juxtaposed compounds are not described in dictionaries and defy analysis. Vice versa, if the hyphen is treated as a punctuation mark they will be analyzed properly but the above examples will be processed erroneously. Therefore graphemic analysis has to be more sophisticated and should consider the hyphen in the both functions.

There is one more problem of morphological tagging which deals with inability of taggers to describe multiwords. This holds true for recognition of collocations that as a rule do not have morphological variances. To ignore them is to distort the linguistic view. These multiwords should be tagged as united lexical units, the so-called compound words or lexemes.

A similar problem is that there are some analytical forms in Russian morphology along with synthetic ones. These forms might be broken-off. How should a tagger treat them?

Blank character function analysis generally is of significant theoretical interest from the point of view of computer modeling of morphology. The traditional approach to this problem is as follows: the blank character is treated as a pure separating signal. Meanwhile such a simplification definitely requires revision.

All phenomena in question should be reflected in the computational model of Russian morphology, which is to reflect language use. In this respect NooJ may be of great use as NooJ dictionaries allow processing both simple and compound words' inflectional morphology in a unified way. That's why building a dictionary of Russian for NooJ is highly desirable.

2.4 Exploring Semantics and NooJ

There is a set of classification and clustering tasks, which provide significant empirical data on lexicon structure, and semantic relations extracted form text corpora. Automatic word clustering (AWC) and document clustering (ADC) as well as word sense disambiguation (WSD) are among those tasks.

AWC shows considerable promise in exposing sets of contextually (and thereby semantically) related words: contextual synonyms/antonyms, hypernyms/ hyponyms, collocates etc. [Pekar, Staab 2003; Pala 2006; Pantel, Lin 2002; Sahlgren 2006]. Data on contextual neighbours of words, on their co-occurrence preferences and on their valency frames extracted from corpora play a crucial role in multilevel text analysis. AWC allows to distinguish words of different semantic classes, so that it can be used in building and enrichment of lexical databases. AWC proves to be quite productive with respect to processing terminological items and domain-restricted texts [Mitrofanova, Panicheva, Savitsky 2007]. AWC is of great importance in document indexing and text summarization which are relevant in ADC procedures [Ferretti, Errecalde, Rosso 2008; Mihalcea, Hassan 2006]. ADC is aimed at finding clusters of texts dealing with similar topics and having similar content. Reliable ADC results provide effective information retrieval and extraction, as well as high quality search [Stein, Meyer zu Eissen 2002]. Experiments with Russian data proved that AWC and ADC data are indispensible in processing and application of various text corpora, they contribute much to adequate domain modeling, as well as ensure development of lexicographic and ontological systems.

WSD alongside with morphological and syntactic disambiguation plays a crucial role in successful corpora development and application. A rich variety of reliable WSD techniques–knowledge-based, corpus-based (statistical), hybrid–have been worked out and tested by now [WSD 2007] (cf. SensEval/SemEval experimental tasks and workshop materials: www.senseval.org, http://nlp.cs.swarthmore.edu/semeval/index.php).

Knowledge-based WSD is performed with the help of semantic information stored in electronic lexicographic modules. Corpus-based WSD implies extraction and statistical processing of word co-occurrence data, which allow to distinguish separate meanings of lexical items in contexts. Hybrid WSD brings into action both lexical resources and corpus analysis.

Experimental data for Russian have shown that it is possible to identify word meanings in contexts using various semantic dictionaries (e.g., RussNet (http://www.project.phil.pu.ru/RussNet/index_ru.shtml), RuTes (http://www.cir.ru/index.jsp), RNC semantic dictionary (http://www.ruscorpora.ru/), etc.), and/or taking into account POS tag distributions [Azarova, Marina 2006], as well as semantic tag distributions and lexical markers [Mitrofanova, Panicheva, Lashevskaya 2008]. Anyway, clustering software should be adjusted for WSD purposes, as WSD requires corpora processing to form clusters of similar contexts.

Due to its functional possibilities and flexibility, NooJ toolkit can be effectively applied in AWC, ADC and WSD procedures conjointly with language-dependent resources and specialized linguistic instruments (corpora, dictionaries, databases, morphosyntactic parsers, taggers, machine learning and pattern recognition modules, etc.).

2.5 Subjectivity Analysis and NooJ

A growing area of research referred to as "subjectivity analysis" implies automatic elaboration of opinions and sentiments expressed in text. Blogs, dialogs, reviews, forums, and even "objective" newspaper articles (which include many opinions and sentiments) are just some of the genres for which accurate identification and interpretation of opinions is critical for full text understanding. Such subjectivity analysis is especially important for service companies because it helps them to evaluate its clients and to improve its activity. Recently NooJ was successfully used for estimation of politeness and satisfaction reflected in dialogs [Ponomareva, Blanco, 2008; Ponomareva, Catena, 2008]. Functional possibilities of NooJ allowed to take into account both lexicographic and stylistic properties of texts under consideration.

It should be mentioned that these characteristics are very important also for constructing the so-called social networks for grouping persons with common interests. Recently this theme became very popular in Internet. And NooJ will find here new applications

Conclusion

In the paper we have presented already existing and possible NooJ-applications for the problems of document clustering and in corpus linguistics. These applications deal with careful indexing and tagging. The next step will be NooJ integration with other modules related with formal document processing as statistical analysis, etc.

References

ALEXANDROV, M., SANCHIS E., ROSSO, P. (2005): Cluster Analysis of Railway Directory Inquire Dialogs. In: “Text, Speech, Dialog”, Springer, LNCS, 3685, pp. 385-392.

ALEXANDROV, M., MAKAGONOV, P. (2007): Introduction to Technique of Clustering. In: Proc. of 3-rd Summer School on Comp. Biology, Brno, pp. 55-80.

AZAROVA, I.V., MARINA, A.S. (2006): Avtomatizirovannaja klassifikacija kontekstov pri podgotovke dannyh dl’a kompjuternogo tezaurusa RussNet. In: Kompjuternaja lingvistika i intellektualnyje tehnologii: Trudy mezhdunarodnoj konferencii «Dialog–2006». Moscow, pp. 13–17 (rus.)

BAEZA-YATES, R., RIBERO-NETO, B. (1999): Modern Information Retrieval. Addison Wesley.

FERRETTI, E., ERRECALDE, M., ROSSO, P. (2008): Does Semantic Information Help in the Text Categorization Task? In: Journal of Intelligent Systems. Vol. 17. № 1–3, pp. 91–107.

LABBE C., LABBE D. (2001): Inter-textual distance and authorship attribution Corneille and Molier. Journ. of Quantitative Linguistics. Vol. 8. № 3, pp. 213–331.

MARUSENKO, M.A. (1990): Atribucija anonimnyh i psevdonimnyh literaturnyh proizvedenij metodami teorii raspoznavanija obrazov. Leningrad.

MIHALCEA, R., HASSAN, S. (2006): Text Summarization for Improved Text Classification. In: Current Issues in Linguistic Theory: Recent Advances in Natural Language Processing. John Benjamins Publishers.

MITROFANOVA, O., PANICHEVA, P., LASHEVSKAYA, O. (2008): Statistical Word Sense Disambiguation in Contexts for Russian Nouns Denoting Physical Objects. In: Text, Speech and Dialogue. LNAI. Springer in press.

MITROFANOVA, O., PANICHEVA, P., SAVITSKY, V. (2007): Automatic Word Clustering in Russian Texts based on Latent Semantic Analysis. In: Computer Treatment of Slavic and East European Languages: 4-th Intern. Seminar. Proceedings. Bratislava, Tribun, pp. 165–175.

MEYER ZU EISSEN, S., STEIN, B., KULIG, M. (2007): Plagiarism Detection without Reference Collections. In: Advances in Data Analysis, Selected Papers from the 30-th Annual Conf. of the German Classication Society (GfKl) Springer, pp. 359–366.

PALA, K. (2006): Word Sketches and Semantic Roles. In: Trudy Mezhdunarodnij Konferencii «Korpusnaja Lingvistika–2006». St. Petersburg, pp. 307–317.

PAN 2007: STEIN, B., KOPPEL, M., STAMATATOS, E. (eds.), In: Proc. of SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (2007).

PANTEL, P., LIN, D. (2002): Discovering Word Senses from Text. In: Proc. of ACM Conference on Knowledge Discovery and Data Mining (KDD-02). Edmonton, Canada, pp. 613–619.

PEKAR, V., STAAB, S. (2003): Word Classification Based on Combined Measures of Distributional and Semantic Similarity. In: Proc. of European Chapter of ACL–03, Research Notes Session. Budapest, pp. 147–150.

PONOMAREVA, N., CATENA, A. (2008): Regression Model for Satisfaction Estimation Trained on Examples. Presented to MICAI-08.

PONOMAREVA, N., BLANCO, X. (2008): Regression Model for Politeness Estimation Trained on Examples. Published in this Proceedings.

RODIONOVA, E.S. (2007): Otbor informativnyh parametrov pri atribucii stihotvornyh pjes Moljera. In: Materialy XXXVI mezhdunarodnoj philologicheskoj konferencii. Vypusk 10. Prikladnaja i matematicheskaja lingvistika. 12–17 marta 2007. St. Petersburg, St. Petersburg State University, Faculty of Philology, pp. 67–74.

—. (2008): Datirovka stihotvornyh pjes, pripisyvajemyh J.-B. Moljeru. In: Materialy XXXVII mezhdunarodnoj philologicheskoj konferencii. Prikladnaja i matematicheskaja lingvistika. 11–15 marta 2008. St. Petersburg, St. Petersburg State University, Faculty of Philology in press.

SAHLGREN, M. (2006): The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces. Ph.D. Dissertation, Department of Linguistics, Stockholm University.

STEIN, B., MEYER ZU EISSEN, S. (2002): Document Categorization with MajorClust. In: Proc. of the 12-th Workshop on Inform. Technology and Systems WITS–02. Barcelona, Spain, pp. 91–96.

STEIN, B., MEYER ZU EISSEN, S. (2003): Automatic Document Categorization: Interpreting the Performance of Clustering Algorithms. In: Proc. 26th German Conf. on Artificial Intelligence, LNCS N 2821, Springer, pp. 254–266.

STEIN, B., MEYER ZU EISSEN, S. (2007): Intrinsic Plagiarism Analysis with Meta Learning. In: Proc.of SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07).

WSD 2007: AGIRRE, E., EDMONDS, Ph. (eds.): Word Sense Disambiguation: Algorithms and Applications. Text, Speech and Language Technology, Vol. 33. Springer (2007)

PORT4NOOJ: PORTUGUESE LINGUISTIC MODULE AND BILINGUAL RESOURCES FOR MACHINE TRANSLATION

ANABELA BARREIRO

Abstract. In this paper we present version 0.1 of *Port4NooJ*, the open source NooJ Portuguese linguistic module, which incorporates a bilingual extension for Portuguese-English (*PT-EN*) machine translation system, (*MT4NooJ*), a work in progress. We first explain the motivation behind this work and then describe the main components of the module, particularly, the electronic dictionaries, the rules which formalize and document Portuguese inflectional and derivational descriptions, and the different types of grammar: morphological, disambiguation, syntactic-semantic, multiword expressions (*MWEs*) and translation grammars. We explain how the different components interact and show the application of these linguistic resources, dictionaries and grammars to text. We present methodology and results driven by the new characteristics of this module.

Key words: Port4NooJ, Portuguese linguistic module, English-Portuguese bilingual resources, morphological grammars, inflectional and derivational descriptions, electronic dictionaries, disambiguation grammars, syntactic-semantic grammars, multiword expressions (*MWEs*), support verb constructions (*SVCs*), grammars for MWEs, paraphrases, translation grammars, machine translation (*MT*)

1. Introduction

NooJ is an established comprehensive linguistic platform which enables the building of language resources such as large coverage dictionaries and multi-purpose grammars (Silberztein, 2004). These language resources can then be applied to single texts or large corpora in real time by the system. Its parsing capabilities include different functions: word or MWEs tagging, location and annotation of morphological, lexical and syntactic-semantic patterns, building concordances, or identification and extraction of semantic units, such as named entities, dates and terminological expressions.

In this paper we present version 0.1 of *Port4NooJ*, the open source PT linguistic module. We describe its main components: the electronic dictionaries, the inflectional and derivational descriptions, and the different types of grammar (morphological, disambiguation, syntactic-semantic, MWE and translation grammars). The module can be downloaded from the NooJ website at http://www.NooJ4nlp.net or directly from http://www.linguateca.pt/Repositorio/Port4NooJ/. Additional resources that we distribute freely to the research community include: a large coverage electronic dictionary (*PT-Dict.dic*); a sample of the morphological rules (inflectional and derivational) (*PT-Sample.flx*); a morphological grammar to process contracted forms (*PT-Contr.nom*); and a syntactic grammar that recognizes and translates dates into English (*PT2EN-Dates.nog*). We also distribute two sample corpora, the PT version of the *Universal Declaration of Human Rights* (*Declaração Universal dos Direitos Humanos*) and the nineteenth century novel *Viagens na Minha Terra* by *Almeida Garret*. These corpora can be used for part of speech (*PoS*) annotation, pattern allocation, semantic units analysis, building concordances, or to extract information from them.

2. Motivation

We developed *Port4NooJ* based on both practical needs and theoretical reasoning. From a practical point of view, we felt the need within the Portuguese research community for large coverage resources that could be used freely by independent researchers (not integrated in research centres with their own resources), so we chose a linguistic environment that we could learn easily and quickly and which would give us the means to build a language description for our specific purposes, that is, the study and formalization of bilingual paraphrases for SVCs to be applied in MT development and evaluation. Just a little experience with *Intex* (Ranchhod et al., 2004) gave us positive results that encouraged us to start building *Port4NooJ* with minimal programming help. We also had experience resulting from our participation in projects such as the development of *Palavroso* (Barreiro et al., 1993) and the development of *Logos* English-Portuguese MT system (Barreiro-Colasuonno, 1999), among others. The skillset required experience in creating dictionaries and grammars, management and integration of open source linguistic resources within large-scale projects. Therefore, when *Logos* Machine Translation (Scott, 2003) became available as open source (beginning of 2006), we decided to use our knowledge of the system to extract and select data and use it for new research, combining the need to build a Portuguese module quickly, so that we could use it for our bilingual paraphrasing, and ultimately develop it into an MT system to translate from PT into EN. We gathered requirements which stated the problem and discovered that the NooJ linguistic environment offered potential and flexibility. For the future, the development offered by other NooJ language modules will make this platform a valuable MT resource, since it has its own multilingual engine; it is already available for a large number of languages with extensive resources, including large-coverage dictionaries and grammars for each individual language. We see this as fertile ground for cross-language studies and development of multilingual applications, including terminological databases. Gathering efforts and sharing resources will be extremely advantageous and, in the long run, NooJ could grow into a successful MT laboratory (*NooJMTLab*). From a theoretical viewpoint, we strongly believe in the framework behind the NooJ linguistic environment which is based on Maurice Gross's legacy, the Lexicon-Grammar Theory (Gross, 1975; 1981). It is a framework which gives primacy to the words in context; i.e., the isolated words have no significance, unless they are integrated in a simple sentence, the elementary unit of meaning. The implications of Gross's legacy regarding translation are also extremely relevant[1]. The challenge found in words has to be undertaken at a multiword level, and the richness of language has to be conveyed in translation. Descriptive representation and computation has to account for language variety and that variety needs to be embodied in the linguistic resources we create.

3. Dictionaries

Port4NooJ contains a large coverage electronic dictionary with nearly 60,000 single word entries (more than a million word forms), represented by their lemmas and classified by part-of-speech (*PoS*), optional

[1] No one better than a close friend could have put it in a more appropriate way: "The translator should pause and mention that, this is what it says, but it could also mean this or that or the other thing, and maybe all at once, or none of the above. When you translate this into French with full consciousness of the words, phrases, and clause structure, and you sense the multiple ambiguities, and feel the rhythm of the words, then you are communicating to your audience the source of the excitement Maurice felt about language." (Extracted from the memorial "*A blank stare at the sunset of the shortest day of the year*", by R. Dougherty, December 21, 2001).

inflectional paradigm (*FLX*), optional derivational paradigm (*DRV*), syntactic-semantic properties, and the corresponding English transfer (*EN*) assigned to each entry, where the "*transfer*" means the disambiguated translation of a word, indispensable for translations of that word in context.

The original dictionary was extracted from *OpenLogos*[2], and then converted into NooJ's format. The conversion involved looking at each numeric representation code of the Syntactic-semantic Abstract Language (*SAL*) ontology of the Logos system and replacing each automatically with understandable SAL mnemonics from the *Logos Tutorial* in the web application, *LearnLogos*[3]. These SAL mnemonics have not been implemented in the Logos system before. Some minor adaptations have been made to these mnemonics. A few sets and subsets are still not converted and most supersets relate to EN and not PT as a source language and they will need to be revised and adapted at a later stage. Once in NooJ format, we inverted it and turned it from an EN-PT to a PT-EN dictionary, cleansed all obsolete entries, separated all multiwords for posterior re-classification and processing, made necessary changes, including modifications and improvements, and added new properties, such as derivation and annotations for predicate nouns and for nominalizations, which we will discuss in further detail later in this section.

All entries are classified as noun, verb, adjective, adverb, determiner, pronoun, preposition, conjunction, or numeric expression. The dictionary contains both invariable and variable word forms. Variable word forms have an inflectional paradigm assigned to them, represented by *FLX*=. For instance, the entry unit *mesa* (EN *table*) inflects according to paradigm class *CASA*, i.e. inflects like the word *casa* (EN *house/home*), and the verb entry *afirmar* (EN *affirm*) conjugates according to class *FALAR*, i.e. inflects like the word *falar* (EN *speak*). Inflectional paradigms are independent standard pattern models (prototypes) based on morphological suffixation rules. These rules cover variation in gender and number (adjectives and nouns), person (verbs and pronouns), tense (verbs), diminutives, augmentatives and superlatives (nouns, adjectives and some adverbs), and nominalizations. Words integrate different hierarchical ontology classes and subclasses, according to their linguistic attributes.

[2] *OpenLogos* is an open source version of the *Logos* machine translation system and is available at http://logos-os.dfki.de/.

[3] SAL represents both the meaning and structure of natural language. See more about SAL in the Logos Tutorial on *OpenLogos*.

Accordingly, syntactic-semantic properties are also provided for each entry. For instance:

1. PT *cão* (EN *dog*) is classified as a *common noun, warm-blooded vertebrate animal, mammal*;
2. PT *vestido* (EN *dress*) is classified as a *concrete noun, clothing, soft thing made of fabric, leather, etc.*;
3. PT *cidade* (EN *city*) is classified as a *common noun, agentive proper name denoting a geographic place, geographical entity, and geographical location*;
4. PT *sair* (EN *leave*) is classified as a *motional intransitive verb*;
5. PT *português* (EN *Portuguese*) is classified as an *adjective, related to a country*;
6. PT *feliz* (EN *happy*) is classified as a *pre-clausal descriptive adjective*;
7. PT *adequadamente* (EN *adequately*) is classified as a *non-locative adverb, manner-type*.

Fig. 1 below shows a small sample of the main dictionary, with representation of all PoS categories, variable and invariable entries (variable entries have specified the inflectional paradigm) and syntactic-semantic properties.

```
mesa,N+FLX=CASA+CO+surf+EN=table
cair,V+FLX=ATRAIR+INMO+IntoType+EN=fall
holandês,A+FLX=INGLÊS+AN+lang+EN=Dutch
actualmente,ADV+FLX=FACILMENTE+TEMP+punc+pres+EN=
nowadays
alguém,PRO+IMPERS+INDEF+EN=somebody
o qual,RELINT+FLX=QUAL+ThatType+EN=which
e,CONJ+JOIN+EN=and
durante,PREP+TEMP+EN=during
cada,DET+IMPERS+INDEF+SG+EN=each
um terço+NUM+frac+EN=one third
```

Fig. 1 General dictionary sample representing all PoS, variable and invariable forms

The properties illustrated in the sample for each entry are the following:

1. *mesa* is classified as a noun (*N*) which inflects like the word *casa* (*FLX=CASA*), where *casa* represents the morphological paradigm for feminine nouns ending in *–a*, plural adding *-s*, with semantic properties defined as *concrete* (*CO*), *functional, bearing surface* (*surf*), corresponding to the EN noun *table*;
2. *cair* is a verb (*V*) which inflects like the verb *atrair* (*FLX=ATRAIR*), where *atrair* represents the morphological paradigm for regular verbs ending in *–air*, vowel change *-i-* > *-í-* in some forms, with syntactic-semantic properties defines as *motional intransitive* (*INMO*), *preposition governance into-type* (*IntoType*), corresponding to the EN verb *fall*;
3. *holandês* is classified as an adjective (*A*) which inflects like the adjective *inglês* (*FLX=INGLÊS*), where *inglês* represents the morphological paradigm for adjectives ending in *–ês*, feminine ending in *-esa*, with syntactic-semantic properties defined as *predicate* (*APred*), *animate* (*AN*) and *language* (*lang*), corresponding to the EN adjective *Dutch*;
4. *actualmente* is an adverb (*ADV*) which inflects like the adverb *facilmente* (*FLX=FACILMENTE*), where *facilmente* represents the morphological paradigm for regular adverbs ending in *–lmente*, superlative in *–íssimamente*, defined as *temporal* (*TEMP*), *punctual* (*punc*), *present* (*pres*), corresponding to the EN adverb *nowadays*;
5. *alguém* is classified as an invariable pronoun (*PRO*), *impersonal* (*IMPERS*), *indefinite* (*INDEF*), corresponding to the EN pronoun *somebody*;
6. *o qual* is classified as a relative and interrogative pronoun (*RELINT*), whose head word *qual* inflects like the pronoun *qual* (*FLX=QUAL*), where *qual* represents the morphological paradigm for pronouns ending in *–al*, plural in *-ais*, with the property *that-type* (*ThatType*), corresponding to the EN pronoun *which*;
7. *e* is classified as a conjunction (*CONJ*), *conjoining* (*JOIN*), corresponding to the EN conjunction *and*;
8. *durante* is classified as a preposition (*PREP*), no inflection, defined as *temporal* (*TEMP*), corresponding to the EN preposition *during*;
9. *cada* is classified as an invariable determiner (*DET*), *impersonal* (*IMPERS*), *indefinite* (*INDEF*), *singular* (*SG*), corresponding to the EN determiner *each*;
10. *um terço* is classified as a numeric expression (*NUM*), *fraction* (*frac*), corresponding to the EN numeric expression *one third*.