XLike Project Language Analysis Services

Xavier Carreras*, **Lluís Padró***, **Lei Zhang**♠, **Achim Rettinger**♠, **Zhixing Li**⋈,
Esteban García-Cuesta◇, **Željko Agić**⋆, **Božo Bekavac**◁, **Blaz Fortuna**†, **Tadej Štajner**†

∗ Universitat Politècnica de Catalunya, Barcelona, Spain. ◇ iSOCO S.A. Madrid, Spain.
◁ University of Zagreb, Zagreb, Croatia. ⋆ University of Potsdam, Germany.
† Jožef Stefan Institute, Ljubljana, Slovenia. ⋈ Tsinghua University, Beijing, China.
♠ Karlsruhe Institute of Technology, Karlsruhe, Germany.

Abstract

This paper presents the linguistic analysis infrastructure developed within the XLike project. The main goal of the implemented tools is to provide a set of functionalities supporting the XLike main objectives: Enabling cross-lingual services for publishers, media monitoring or developing new business intelligence applications. The services cover seven major and minor languages: English, German, Spanish, Chinese, Catalan, Slovenian, and Croatian. These analyzers are provided as web services following a lightweigth SOA architecture approach, and they are publically accessible and shared through META-SHARE.[1]

1 Introduction

Project XLike[2] goal is to develop technology able to gather documents in a variety of languages and genres (news, blogs, tweets, etc.) and to extract language-independent knowledge from them, in order to provide new and better services to publishers, media monitoring, and business intelligence. Thus, project use cases are provided by STA (Slovenian Press Agency) and Bloomberg, as well as New York Times as an associated partner.

Research partners in the project are Jožef Stefan Institute (JSI), Karlsruhe Institute of Technology (KIT), Universitat Politècnica de Catalunya (UPC), University of Zagreb (UZG), and Tsinghua University (THU). The Spanish company iSOCO is in charge of integration of all components developed in the project.

This paper deals with the language technology developed within the project XLike to convert input documents into a language-independent representation that afterwards enables knowledge aggregation.

To achieve this goal, a bench of linguistic processing pipelines is devised as the first step in the document processing flow. Then, a cross-lingual semantic annotation method, based on Wikipedia and Linked Open Data (LOD), is applied. The semantic annotation stage enriches the linguistic anaylsis with links to knowledge bases for different languages, or links to language independent representations.

2 Linguistic Analyzers

Apart from basic state-of-the-art tokenizers, lemmatizers, PoS/MSD taggers, and NE recognizers, each pipeline requires deeper processors able to build the target language-independent semantic representantion. For that, we rely on three steps: dependency parsing, semantic role labeling and word sense disambiguation. These three processes, combined with multilingual ontological resouces such as different WordNets and PredicateMatrix (López de la Calle et al., 2014), a lexical semantics resource combining WordNet, FrameNet, and VerbNet, are the key to the construction of our semantic representation.

2.1 Dependency Parsing

We use graph-based methods for dependency parsing, namely, **MSTParser**[3] (McDonald et al., 2005) is used for Chinese and Croatian, and **Treeler**[4] is used for the other languages. Treeler is a library developed by the UPC team that implements several statistical methods for tagging and parsing.

We use these tools in order to train dependency parsers for all XLike languages using standard available treebanks.

[1] *accessible and shared* here means that the services are publicly callable, not that the code is open-source. http://www.meta-share.eu

[2] http://www.xlike.org

[3] http://sourceforge.net/projects/mstparser

[4] http://treeler.lsi.upc.edu

Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 9–12,
Gothenburg, Sweden, April 26-30 2014.

2.2 Semantic Role Labeling

As with syntactic parsing, we are developing SRL methods with the Treeler library. In order to train models, we will use the treebanks made available by the CoNLL-2009 shared task, which provided data annotated with predicate-argument relations for English, Spanish, Catalan, German and Chinese. No treebank annotated with semantic roles exists for Slovene or Croatian. A prototype of SRL has been integrated in all pipelines (except the Slovene and Croatian pipelines). The method implemented follows a pipeline architecture described in (Lluís et al., 2013).

2.3 Word Sense Disambiguation

Word sense disambiguation is performed for all languages with a publicly available WordNet. This includes all languages in the project except Chinese. The goal of WSD is to map specific languages to a common semantic space, in this case, WN synsets. Thanks to existing connections between WN and other resources, SUMO and OpenCYC sense codes are also output when available.

Thanks to PredicateMatrix, the obtained concepts can be projected to FrameNet, achieving a normalization of the semantic roles produced by the SRL (which are treebank-dependent, and thus, not the same for all languages). The used WSD engine is the UKB (Agirre and Soroa, 2009) implementation provided by FreeLing (Padró and Stanilovsky, 2012).

2.4 Frame Extraction

The final step is to convert all the gathered linguistic information into a semantic representation. Our method is based on the notion of frame: a semantic frame is a schematic representation of a situation involving various participants. In a frame, each participant plays a role. There is a direct correspondence between roles in a frame and semantic roles; namely, frames correspond to predicates, and participants correspond to the arguments of the predicate. We distinguish three types of participants: entities, words, and frames.

Entities are nodes in the graph connected to real-world entities as described in Section 3. Words are common words or concepts, linked to general ontologies such as WordNet. Frames correspond to events or predicates described in the document. Figure 1 shows an example sentence, the extracted frames and their arguments.

It is important to note that frames are a more general representation than SVO-triples. While SVO-triples represent a binary relation between two participants, frames can represent n-ary relations (e.g. predicates with more than two arguments, or with adjuncts). Frames also allow representing the sentences where one of the arguments is in turn a frame (as is the case with *plan to make* in the example).

Finally, although frames are extracted at sentence level, the resulting graphs are aggregated in a single semantic graph representing the whole document via a very simple coreference resolution based on detecting named entity aliases and repetitions of common nouns. Future improvements include using an state-of-the-art coreference resolution module for languages where it is available.

3 Cross-lingual Semantic Annotation

This step adds further semantic annotations on top of the results obtained by linguistic processing. All XLike languages are covered. The goal is to map word phrases in different languages into the same semantic interlingua, which consists of resources specified in knowledge bases such as Wikipedia and Linked Open Data (LOD) sources. Cross-lingual semantic annotation is performed in two stages: (1) first, candidate concepts in the knowledge base are linked to the linguistic resources based on a newly developed cross-lingual linked data lexica, called xLiD-Lexica, (2) next the candidate concepts get disambiguated based on the personalized PageRank algorithm by utilizing the structure of information contained in the knowledge base.

The xLiD-Lexica is stored in RDF format and contains about 300 million triples of cross-lingual groundings. It is extracted from Wikipedia dumps of July 2013 in English, German, Spanish, Catalan, Slovenian and Chinese, and based on the canonicalized datasets of DBpedia 3.8 containing triples extracted from the respective Wikipedia whose subject and object resource have an equivalent English article.

4 Web Service Architecture Approach

The different language functionalities are implemented following the service oriented architecture (SOA) approach defined in the project XLike. Therefore all the pipelines (one for each language) have been implemented as web services and may

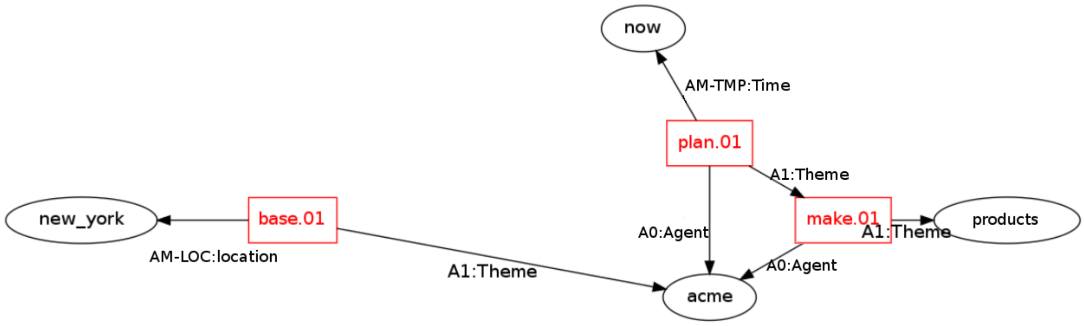

Figure 1: Graphical representation of frames in the sentence *Acme, based in New York, now plans to make computer and electronic products.*

be requested to produce different levels of analysis (e.g. tokenization, lemmatization, NERC, parsing, relation extraction). This approach is very appealing due to the fact that it allows to treat every language independently and execute the whole language analysis process at different threads or computers allowing an easier parallelization (e.g. using external high perfomance platforms such as Amazon Elastic Compute Cloud EC2[5]) as needed. Furthermore it also provides independent development lifecycles for each language which is crucial in this type of research projects. Recall that these web services can be deployed locally or remotely, maintaining the option of using them in a stand-alone configuration.

The main structure for each one of the pipelines is described below:

- **Spanish**, **English**, and **Catalan**: all modules are based on FreeLing (Padró and Stanilovsky, 2012) and Treeler.
- **German**: German shallow processing is based on OpenNLP[6], Stanford POS tagger and NE extractor (Toutanova et al., 2003; Finkel et al., 2005). Dependency parsing, semantic role labeling, word sense disambiguation, and SRL-based frame extraction are based on FreeLing and Treeler.
- **Slovene**: Slovene shallow processing is provided by JSI Enrycher[7] (Štajner et al., 2010), which consists of the Obeliks morphosyntactic analysis library (Grčar et al., 2012), the LemmaGen lemmatizer (Juršič et al., 2010) and a CRF-based entity extractor (Štajner et al., 2012). Dependency parsing, word sense disambiguation are based on FreeLing and Treeler. Frame extraction is rule-based since no SRL corpus is available for Slovene.
- **Croatian**: Croatian shallow processing is based on proprietary tokenizer, POS/MSD-tagging and lemmatisaton system (Agić et al., 2008), NERC system (Bekavac and Tadić, 2007) and dependency parser (Agić, 2012). Word sense disambiguation is based on FreeLing. Frame extraction is rule-based since no SRL corpus is available for Croatian.
- **Chinese**: Chinese shallow and deep processing is based on a word segmentation component ICTCLAS[8] and a semantic dependency parser trained on CSDN corpus. Then, rule-based frame extraction is performed (no SRL corpus nor WordNet are available for Chinese).

Each language analysis service is able to process thousands of words per second when performing shallow analysis (up to NE recognition), and hundreds of words per second when producing the semantic representation based on full analysis. Moreover, the web service architecture enables the same server to run a different thread for each client, thus taking advantage of multiprocessor capabilities.

The components of the cross-lingual semantic annotation stage are:

- **xLiD-Lexica**: The cross-lingual groundings in xLiD-Lexica are translated into RDF data and are accessible through a SPARQL endpoint, based on OpenLink Virtuoso[9] as the back-end database engine.

[5] http://aws.amazon.com/ec2/
[6] http://opennlp.apache.org
[7] http://enrycher.ijs.si
[8] http://ictclas.org/
[9] http://virtuoso.openlinksw.com/

- **Semantic Annotation**: The cross-lingual semantic annotation service is based on the xLiD-Lexica for entity mention recognition and the JUNG Framework[10] for graph-based disambiguation.

5 Conclusion

We presented the web service based architecture used in XLike FP7 project to linguistically analyze large amounts of documents in seven different languages. The analysis pipelines perform basic processing as tokenization, PoS-tagging, and named entity extraction, as well as deeper analysis such as dependency parsing, word sense disambiguation, and semantic role labelling. The result of these linguistic analyzers is a semantic graph capturing the main events described in the document and their core participants.

On top of that, the cross-lingual semantic annotation component links the resulting linguistic resources in one language to resources in a knowledge bases in any other language or to language independent representations. This semantic representation is later used in XLike for document mining purposes such as enabling cross-lingual services for publishers, media monitoring or developing new business intelligence applications.

The described analysis services are currently available via META-SHARE as callable RESTful services.

Acknowledgments

This work was funded by the European Union through project XLike (FP7-ICT-2011-288342).

References

Željko Agić, Marko Tadić, and Zdravko Dovedan. 2008. Improving part-of-speech tagging accuracy for Croatian by morphological analysis. *Informatica*, 32(4):445–451.

Željko Agić. 2012. K-best spanning tree dependency parsing with verb valency lexicon reranking. In *Proceedings of COLING 2012: Posters*, pages 1–12, Mumbai, India, December. The COLING 2012 Organizing Committee.

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

[10]Java Universal Network/Graph Framework http://jung.sourceforge.net/

Božo Bekavac and Marko Tadić. 2007. Implementation of Croatian NERC system. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing (BSNLP2007), Special Theme: Information Extraction and Enabling Technologies*, pages 11–18. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*, pages 363–370.

Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In *Zbornik Osme konference Jezikovne tehnologije*, Ljubljana, Slovenia.

Matjaz Juršič, Igor Mozetič, Tomaz Erjavec, and Nada Lavrač. 2010. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9):1190–1214.

Xavier Lluís, Xavier Carreras, and Lluís Màrquez. 2013. Joint arc-factored parsing of syntactic and semantic dependencies. *Transactions of the Association for Computational Linguistics*, 1:219–230.

Maddalen López de la Calle, Egoitz Laparra, and German Rigau. 2014. First steps towards a predicate matrix. In *Proceedings of the Global WordNet Conference (GWC 2014)*, Tartu, Estonia, January. GWA.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 91–98, Ann Arbor, Michigan, June.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Tadej Štajner, Delia Rusu, Lorand Dali, Blaž Fortuna, Dunja Mladenić, and Marko Grobelnik. 2010. A service oriented framework for natural language text enrichment. *Informatica*, 34(3):307–313.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Lin- guistics on Human Language Technology (NAACL'03)*.

Tadej Štajner, Tomaž Erjavec, and Simon Krek. 2012. Razpoznavanje imenskih entitet v slovenskem besedilu. In *In Proceedings of 15th Internation Multiconference on Information Society - Jezikovne Tehnologije*, Ljubljana, Slovenia.