Langforia: Language Pipelines for Annotating Large Collections of Documents

Marcus Klang
Lund University
Department of Computer Science
Lund, Sweden
Marcus.Klang@cs.lth.se

Pierre Nugues
Lund University
Department of Computer Science
Lund, Sweden
Pierre.Nugues@cs.lth.se

Abstract

In this paper, we describe **Langforia**, a multilingual processing pipeline to annotate texts with multiple layers: formatting, parts of speech, named entities, dependencies, semantic roles, and entity links. Langforia works as a web service, where the server hosts the language processing components and the client, the input and result visualization. To annotate a text or a Wikipedia page, the user chooses an NLP pipeline and enters the text or the name of the Wikipedia page in the input field of the interface. Once processed, the results are returned to the client, where the user can select the annotation layers s/he wants to visualize.

We designed Langforia with a specific focus for Wikipedia, although it can process any type of text. Wikipedia has become an essential encyclopedic corpus used in many NLP projects. However, processing articles and visualizing the annotations are nontrivial tasks that require dealing with multiple markup variants, encodings issues, and tool incompatibilities across the language versions. This motivated the development of a new architecture.

A demonstration of Langforia is available for six languages: English, French, German, Spanish, Russian, and Swedish at `http://vilde.cs.lth.se:9000/` as well as a web API: `http://vilde.cs.lth.se:9000/api`. Langforia is also provided as a standalone library and is compatible with cluster computing.

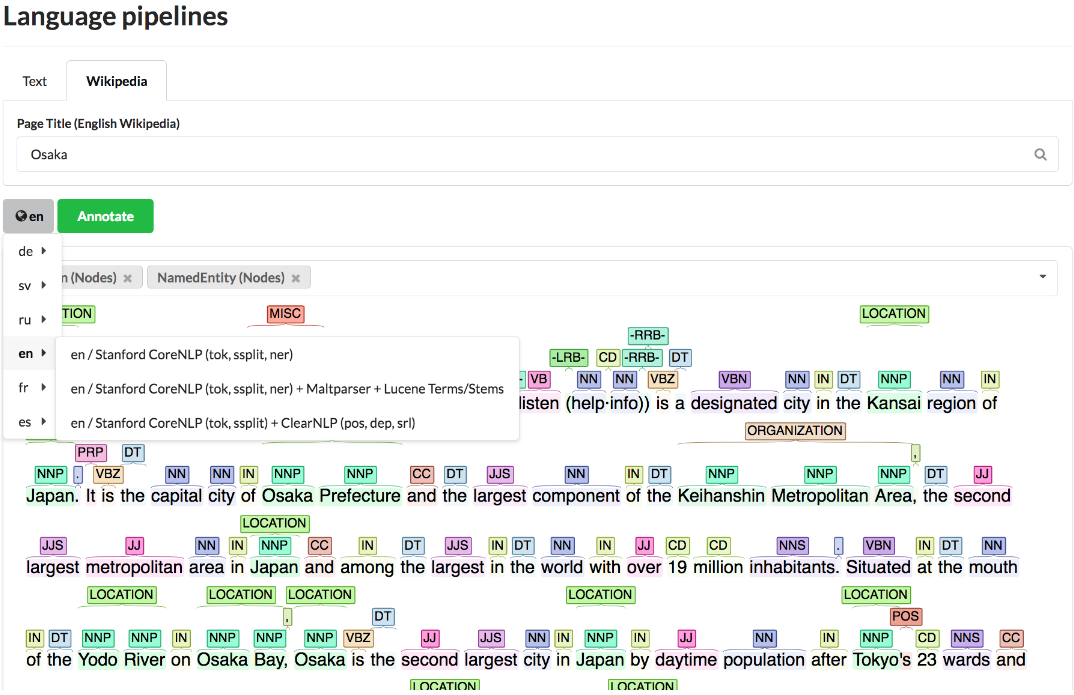
1 The Demonstration

Langforia is a multilingual annotation and visualization platform available as a web service and as a standalone library. Figure 1 shows the interface, where the user chooses the language and tool chain s/he wants to use from the drop-down menu to the left. Depending on the language and the availability of components, the annotations can range from tokenization to dependency parsing, semantic role labeling, and entity linking. The user then either enters a text or writes the name of a Wikipedia page and presses the "Annotate" button. If the document to analyze is a raw text, it is sent directly to the server; if it is a Wikipedia page name, the client first fetches the HTML content of this page from `https://www.wikipedia.org/` and then sends it to the Langforia server. Figure 2, left part, shows the resulting annotations for the *Osaka* article from the Swedish Wikipedia for three layers, tokens, named entities, and dependency relations, while the right part of the figure shows the entity linking results.

2 Motivation and Significance

We designed Langforia with a specific focus for Wikipedia, although the pipeline can process raw text. Wikipedia has become an essential encyclopedic corpus used in many NLP projects. In translation (Smith et al., 2010), semantic networks (Navigli and Ponzetto, 2010), named entity linking (Mihalcea and Csomai, 2007), information extraction, or question answering (Ferrucci, 2012), Wikipedia offers a multilingual coverage and an article diversity that are unequalled. However, processing articles are nontrivial tasks that require dealing with multiple markup variants, encodings issues, tool incompatibilities

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations,
pages 74–78, Osaka, Japan, December 11-17 2016.



Osaka

en / Stanford CoreNLP (tok, ssplit, ner)

en / Stanford CoreNLP (tok, ssplit, ner) + Maltparser + Lucene Terms/Stems

en / Stanford CoreNLP (tok, ssplit) + ClearNLP (pos, dep, srl)

listen (help·info)) is a designated city in the Kansai region of Japan. It is the capital city of Osaka Prefecture and the largest component of the Keihanshin Metropolitan Area, the second largest metropolitan area in Japan and among the largest in the world with over 19 million inhabitants. Situated at the mouth of the Yodo River on Osaka Bay, Osaka is the second largest city in Japan by daytime population after Tokyo's 23 wards and

Figure 1: The Langforia interface. The upper part of the figure shows the input box, where the user either selects a Wikipedia page (Wikipedia tab), here the article on Osaka in English, or enters a raw text (Text tab); The center part shows the language selection menu with six languages (de, en, es, fr, ru, sv), here English, and submenus to choose the toolchain (three toolchains for English); and the Annotate button; The lower part shows the annotated text, where the annotation layers are selectable from a drop down menu in the block just above (black triangle to the right), here the tokens and named entities

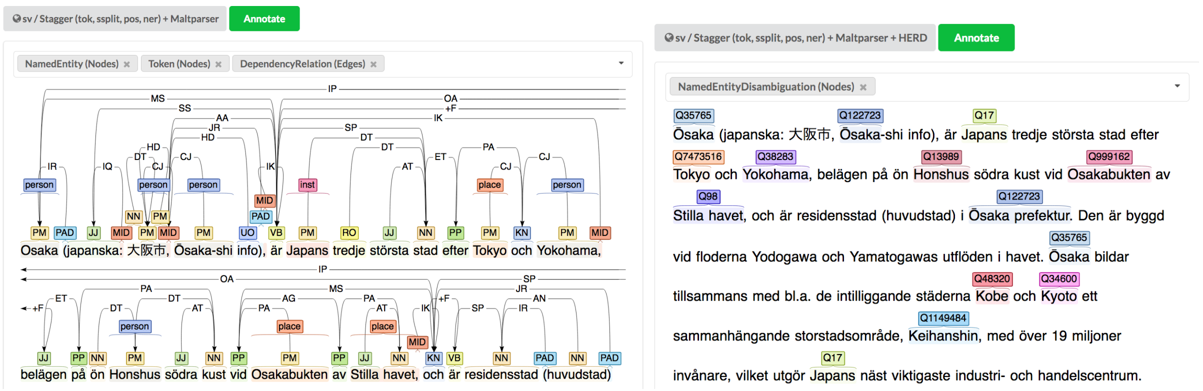
across the language versions and significant processing capacities. In addition, the scale and heterogeneity of the Wikipedia collection makes it relatively difficult to do experimentations on the whole corpus. These experimentations are rendered even more complex as, to the best of our knowledge, there is no available tool to visualize easily annotation results from different processing pipelines.

Langforia builds on a document model (Klang and Nugues, 2016) that stores the linguistic annotations and enables the pipeline to abstract the components across the languages and tools. This model consists of layers, where each layer is a sequence of ranges describing a specific annotation, for instance the parts of speech or the syntactic dependencies. It provides a format common to all the pipelines that makes them insensitive to the input/output features of a tool.

The list of annotated layers varies depending on the tool availability for a specific language. The layers common to all the versions are compatible with the Wikipedia markup: They include the text, paragraphs, text styles, links, and page sections. Using this document model as input, we created a client visualizer that let users interactively visualize the annotations. Beyond the demonstration, Langforia is available in the form of a library that provides a uniform way to process multilingual Wikipedia dumps and output the results in a universal document model. This could benefit all the projects that use Wikipedia as a corpus.

3 System Architecture

Langforia consists of three parts: A set of language processing components assembled as tool chains; a multilayer document model (MLDM) library; and a visualizer.


sv / Stagger (tok, ssplit, pos, ner) + Maltparser
Annotate
NamedEntity (Nodes)
Token (Nodes)
DependencyRelation (Edges)
person
inst
place
sv / Stagger (tok, ssplit, pos, ner) + Maltparser + HERD
Annotate
NamedEntityDisambiguation (Nodes)
Q35765
Q122723
Q17
Q7473516
Q38283
Q13989
Q999162
Q98
Q122723
Q35765
Q48320
Q34600
Q1149484
Q17

Osaka (japanska: 大阪市, Ōsaka-shi info), är Japans tredje största stad efter Tokyo och Yokohama,
belägen på ön Honshus södra kust vid Osakabukten av Stilla havet, och är residensstad (huvudstad)

Ōsaka (japanska: 大阪市, Ōsaka-shi info), är Japans tredje största stad efter Tokyo och Yokohama, belägen på ön Honshus södra kust vid Osakabukten av Stilla havet, och är residensstad (huvudstad) i Ōsaka prefektur. Den är byggd vid floderna Yodogawa och Yamatogawas utflöden i havet. Ōsaka bildar tillsammans med bl.a. de intilliggande städerna Kobe och Kyoto ett sammanhängande storstadsområde, Keihanshin, med över 19 miljoner invånare, vilket utgör Japans näst viktigaste industri- och handelscentrum.

Figure 2: Left part: Visualization of three layers: Tokens, named entities, and dependency relations from the *Osaka* page in Swedish; right part: Visualization of named entity linking with Wikidata identifiers

3.1 Tool Chains

We annotate Wikipedia HTML pages into MLDM records using an annotation pipeline: a sequence of processing components. The first step converts the HTML documents into DOM trees using jsoup[1]. The second step extracts the original page structure, text styles, links, lists, and tables. We then resolve the links to unique Wikidata identifiers. Wikidata is an entity database[2], part of Wikimedia, which assigns unique identifiers across all the language editions of Wikipedia. The city of Osaka, for instance, has the unique id: Q35765 that enables the system to retrieve the article pages in English, French, Swedish, or Russian. We keep the original links occurring in the Wikipedia pages and we resolve them using Wikidata identifiers, when they exist, or to normalized page names as a fall back. These steps are common to all the language editions we process. If the input is plain text, we skip these steps.

The annotation tool chains are specific to the languages. We abstracted these chains so that they are instances of a generic annotator. For English, Spanish, and German, we use CoreNLP (Manning et al., 2014) or ClearNLP (Choi, 2012). For French, we use CoreNLP for tokenizing the text and MATE for parsing (Björkelund et al., 2010). For Swedish, we use Stagger (Östling, 2013) and MaltParser (Nivre et al., 2006). For Russian, only the tokenization is available for now. We also link mentions of named entities and concepts to unique Wikidata identifiers. To carry this out, we reimplemented a variant of TAGME (Ferragina and Scaiella, 2010).

3.2 The Document Model

The MLDM library[3] (Klang and Nugues, 2016) defines a model for storing, querying, and extracting hypertextual information common to many NLP tasks in a standalone package. We designed this model so that it could store the original Wikipedia markup, as well as the subsequent linguistic annotations: Part-of-speech tagging, coreference resolution, named entity recognition and linking, dependency parsing, semantic role labeling, etc.

The model consists of multiple layers, where each layer is dedicated to a specific type of annotation. The annotations are encoded in the form of graph nodes, where a node represents a piece of data: a token, a sentence, a named entity, etc., delimited by ranges. These nodes are possibly connected by edges as in dependency graphs. This data structure used is similar to a property graph.

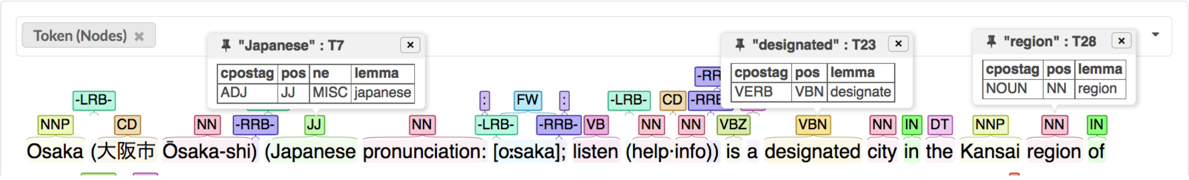
3.3 Visualization

The interactive visualization tool enables the user to examine the results. We designed it so that it could handle large documents with more than 10,000 tokens with a fast rendering of the annotations and allow cross sentence annotations, such as for paragraphs and sections. The layers are selectable from a dropdown menu and the supported visualizations are the ranges and relationships between them.

[1] `http://jsoup.org/`

[2] `http://www.wikidata.org`

[3] `https://github.com/marcusklang/docforia`



Osaka (大阪市 Ōsaka-shi) (Japanese pronunciation: [oːsaka]; listen (help·info)) is a designated city in the Kansai region of

"Japanese" : T7

cpostag	pos	ne	lemma
ADJ	JJ	MISC	japanese

"designated" : T23

cpostag	pos	lemma
VERB	VBN	designate

"region" : T28

cpostag	pos	lemma
NOUN	NN	region

Figure 3: The properties attached to the words *Japanese*, *designated*, and *region*, in the form of tooltips

In Fig. 3, we selected the token layer that by default displays the parts of speech of the words. If we hover over the words, the visualizer shows the properties attached to a word in CoNLL-like format in a tooltip that the user can fix, move, and discard. Figure 3 shows the properties of the words: *Japanese*, *designated*, and *region*. Finally, we estimated the rendering speed (time to interactive use) on 30,000 annotations (tokens) with Intel Core i7, 2.3 GHz, with 16 GB RAM running a Chrome browser and we obtained the figure of 7.7s seconds, i.e. 3,800 annotations per second.

4 Related Work

The UIMA project (Ferrucci and Lally, 2004) provides an infrastructure to store unstructured documents. In contrast, the MLDM library and Langforia emphasize on simplicity, portability, ease of integration, minimal dependencies, and efficiency. Other toolchains include CoreNLP (Manning et al., 2014). However, CoreNLP cannot process the Wikipedia markup or easily integrate external tools. In addition, CoreNLP does not provide a storage model and its data structures are primarily meant to extend its functionalities. In contrast to CoreNLP, Langforia builds on Docforia that provides dynamic and typed annotations as well as multiple sublayers such as gold and predicted. Finally, CoreNLP does not provide a query API for its data structures.

The Langforia visualization tool is similar to the brat[4] components (Stenetorp et al., 2012) for the text visualization. Brat produces good visual results and has support for multiple layers of information. However, to the best of our knowledge, it lacks tooltip support in the embeddable version and it does not handle line-wrapped annotations well. In addition, it revealed too slow to render a large number of annotations in the documents we tested.

5 Conclusion and Future work

We described Langforia, a multilingual tool for processing text and visualizing annotations. Langforia builds on a multilayer document model (MLDM), structured in the form of a graph and unified tool chains. It enables a user to easily access the results of multilingual annotations and through its API to process large collections of text. Using it, we built a tabulated version of Wikipedia (Klang and Nugues, 2016) that can be queried using a SQL-like language. When applied to Wikipedia, MLDM links the different versions through an extensive use of URI indices and Wikidata Q-numbers.

6 Availability

The Langforia demonstration is accessible at: `http://vilde.cs.lth.se:9000/` and the web API at: `http://vilde.cs.lth.se:9000/api`. The source code is available from github at: `https://github.com/marcusklang/`.

Acknowledgments

This research was supported by Vetenskapsrådet, the Swedish research council, under the *Det digitaliserade samhället* program.

[4] `http://brat.nlplab.org/`

References

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstration Volume*, pages 33–36, Beijing, August 23-27.

Jinho D. Choi. 2012. *Optimization of Natural Language Processing Components for Robustness and Scalability*. Ph.D. thesis, University of Colorado at Boulder, Boulder, CO, USA. AAI3549172.

Paolo Ferragina and Ugo Scaiella. 2010. Fast and accurate annotation of short texts with wikipedia pages. In *Proceedings of CIKM'10*, Toronto.

David Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, September.

David Angelo Ferrucci. 2012. Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3.4):1:1 –1:15, May-June.

Marcus Klang and Pierre Nugues. 2016. Wikiparq: A tabulated Wikipedia resource using the Parquet format. In *Proceedings of LREC 2016*, pages 4141–4148, Portorož, Slovenia.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on CIKM*, CIKM '07, pages 233–242, Lisbon, Portugal.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the ACL*, pages 216–225, Uppsala.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC-2006*, pages 2216–2219.

Robert Östling. 2013. Stagger: an open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology*, 3:1–18.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 403–411.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics.