Features for Generic Corpus Querying

Thomas Eckart, Christoph Kuras, Uwe Quasthoff
Natural Language Processing Group, University of Leipzig
Augustusplatz 10, 04109 Leipzig, Germany
Email: {teckart, ckuras, quasthoff}@informatik.uni-leipzig.de

Abstract

The availability of large corpora for more and more languages enforces generic querying and standard interfaces. This development is especially relevant in the context of integrated research environments like CLARIN or DARIAH. The paper focuses on several applications and implementation details on the basis of a unified corpus format, a unique POS tag set, and prepared data for word similarities. All described data or applications are already or will be in the near future accessible via well-documented RESTful Web services. The target group are all kinds of interested persons with varying level of experience in programming or corpus query languages.

Keywords: corpus, corpus query, Web services

1. Introduction

A corpus is called generic if it is designed to fit for multiple purposes. This usually implies a comparably large size, high quality and a certain kind of added value like annotation or additional pre-calculated data. The Leipzig Corpora Collection provides generic corpora for more than 200 languages. In 2015 enhanced functionality were developed and new features were added to simplify the usage of the LCC data as a generic source for annotated text material. The data can be queried using a newly developed Web frontend[1] or directly via Web services[2].

Some of the feature values described below depend on the quality of the corpus. For instance, duplicates (and near duplicates) in the corpus generate unnatural significant word co-occurrences (see sect. 2.3). So, only extensive preprocessing with emphasis on corpus quality will generate reliable data with general acceptance.

2. Feature Overview

The following list of features can be considered as a standard set of corpus queries. These features have a different range of application like general linguistics, lexicography, information extraction, and ontology learning. Recipients are either human users or other algorithms. The features are designed to have quick response time which allows many queries. Moreover, they should serve multiple interests and the output should be usable for all kinds of recipients.

2.1. Word frequency and distribution

The following information is available for all words in a corpus and also for a pre-selected list of multiword units (MWUs). For a certain language, this list of MWUs contains all article titles from Wikipedia in this language and possibly more data like additional person names, technical terms and phraseology. In the following, the term word is also used for these MWUs.

The word frequency is available not only for the whole corpus. In the case of news corpora where a corpus production is performed on a yearly basis there is a frequency distribution over time for every word. Moreover, there is a distribution w.r.t. sources which shows whether a word is of general use or belongs to a subset like a subject area

2.2. Part-of-Speech Tags

A set of different POS-taggers is used to provide POS-tags for as many languages as possible. If a tagger for the language of a corpus is available, POS-tags will be created. Currently this is the case for 34 languages. Despite the tagged sentences there is frequency information for each combination of word and tag available. Some taggers also support the output of the lemma form of a word, which will also be part of the corpus if available.

The corpora used here include several languages. Due to the different models used for each language the tagsets vary which makes the results difficult to compare. To address this problem the approach of Universal Dependencies (Petrov et al, 2011) is used. For each tagset a mapping is created which transforms the original tags to UD17, a set of 17 universal POS-Tags[3] (see Table 1). In order to reduce manual work for creation of the mappings,

1 Available at http://corpora.informatik.uni-leipzig.de/
2 http://wortschatzwebservices.informatik.uni-leipzig.de
3 http://universaldependencies.github.io/docs/u/pos/

in the near future Lingua Interset[4] will be used for conversion from the original tagsets. This useful tool includes import and export functionality for a high number of existing tagsets.

Tag	*Description*
ADJ	adjective
ADP	adposition
ADV	adverb
AUX	auxiliary verb
CONJ	coordinating conjunction
DET	determiner
INTJ	interjection
NOUN	noun
NUM	numeral
PART	particle
PRON	pronoun
PROPN	proper noun
SCONJ	punctuation
SYM	subordinating conjunction
VERB	verb
X	other

Table 1: Universal tag set UD17.

This allows the user not only to compare tagging results of different models for the same language but to compare sentence patterns interlingually.
Figure 1 shows the distribution of UD17 POS tags for different languages based on complete corpora. This visualization demonstrates the similarity both of the behaviour of the different underlying POS taggers and the corresponding languages.

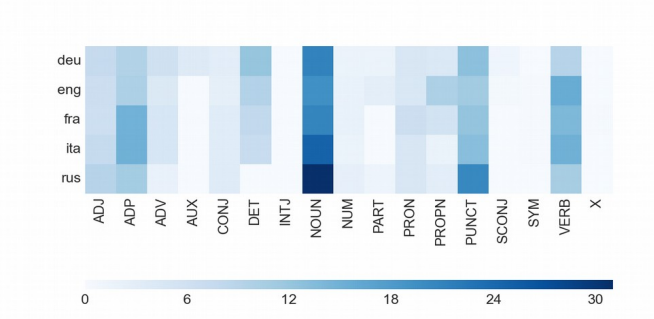

Figure 1: Distribution of UD17 POS tags for different languages.

Moreover, the distribution of ambiguous POS tags (so-called ambitags) in the different languages can be compared. Different POS tags can be given to a word appearing in different contexts. This can either be the consequence of a linguistic ambiguity or of a tagger error. Therefore the distribution of ambitags can show both language similarities and tagger problems.
Figure 2 shows the most frequent ambitags for the top-100.000 words for different languages. For every pair (word, POS tag) a minimum frequency of one per million tokens was applied.

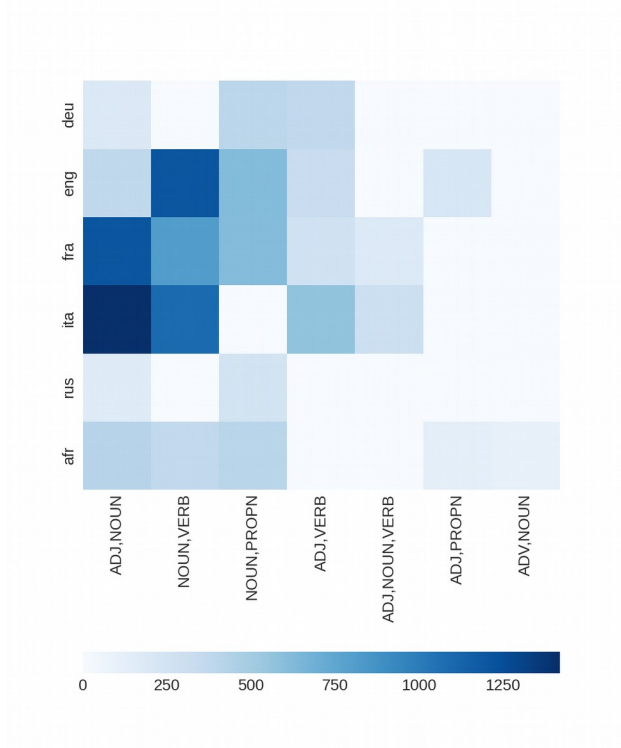

Figure 2: Most frequent ambitags for different languages.

2.3. Word co-occurrences

Word co-occurrences are pairs of words which appear significantly often together. Especially interesting is the joint occurrence as next neighbor or within a larger window like a sentence. Word co-occurrences often represent interesting relations (Heyer et al. 2001). If a corpus is large enough, these relations are of general interest and not related to the exact composition of the underlying corpus.
For each pair of word co-occurrences (as next neighbors or sentence-based) both the log-Likelihood significance measure (Dunning, 1993) and the frequency are given. As described above in the case of words, these numbers are available to generate a distribution over time and for different sources.

2.4. Word similarity

The following two kinds of word similarity are pre-calculated: String similarity of words using a fast Levenshtein algorithm is available for all words and distances up to two. This helps for spelling variants

4 https://metacpan.org/pod/Lingua::Interset

(Gorbachev, Gorbachov, Gorbatchev, Gorbechov, ...) and even (if applied recursively) can show the most important morphological transformations. Table 2 shows the most frequent affixes with a maximal length of two characters for the 100.000 most frequent German words.

Prefix	*Frequency*	*Suffix*	*Frequency*
ge	651	-n	8065
be-	580	-en	5845
un-	565	-e	4757
An-	446	-s	3232
ab-	433	-r	1861
er-	322	-er	1856
zu-	192	-es	1186

Table 2: Frequent word affixes in a German corpus.

A semantic word similarity is based on context similarity. For two words, the numbers of joint word co-occurrences is counted. Table 3 gives examples for English colours, German weekdays, and Russian months.

Language	*Start word*	*Most similar words*
English	yellow	red, blue, white, green, pink, black, orange, purple, ...
German	Montag *[Monday]*	Dienstag, Donnerstag, Mittwoch, Freitag, Samstag, Sonntag
Russian	января *[January, gen.]*	Декабря, мая, марта, апреля, ноября, сентября, октября, февраля, июля, июня, августа

Table 3: Similar words ordered by decreasing similarity.

All these similarity data are time-consuming to calculate and therefore pre-calculated.

2.5. Sample sentences with GDEX ranking

Sample sentences are provided for single words and word co-occurrences. For human users, “nice” sentences are important. In (Kilgarriff et al. 2008), the GDEX algorithm was presented to select nice sentences for a given word. These criteria were developed further to a global ordering imposed on all sentences. Sample sentences are then selected according to this global ordering. The following criteria impose penalties on every sentence, and the sentences with the least combined penalty are used as examples.

- Sentence length in characters: Sentences longer than 100 characters are penalized with increasing length.
- Special characters (except the standard punctuation marks ,.!? are penalized).
- An odd number of quotation marks are strongly penalized.
- Every digit is penalized.
- Every pair of consecutive uppercase letters is penalized.
- The number of stop words in the sentence should be about 40%.
- The average word rank should be about 8000.
- The rarest word in the sentence should have a rank of about 32000.
- The average word length should be slightly less than average.

As result, the penalties are distributed for different languages as depicted in Figure 3. Here, the percentage of sentences in the corpus with a (rounded) penalty value is illustrated.

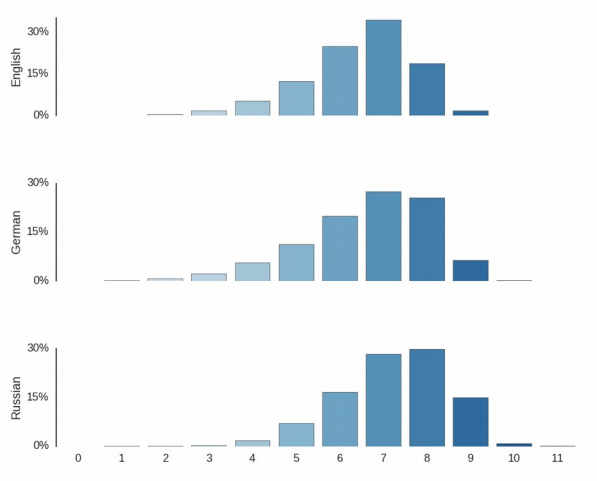

Figure 3: GDEX-based penalties for several corpora in different languages (Percentage of sentences in a corpus for a specific (rounded) penalty value).

2.6. Universal POS tags and Nosketch-Engine

The NoSketch-Engine (Rychlý, 2007) is a corpus-management system suitable for working with generic corpora. It is an open source project based on the commercial Sketch Engine (Kilgarriff, 2014) service. With the NoSketch-Engine the user is able to explore large corpora quickly using basic string search or the more advanced Corpus Query Language (CQL). Additionally it provides access to word frequency lists, collocations and other useful features. One of the key features concerning generic corpora is the possibility to access and query any annotation contained in the corpus making it a flexible tool for many different scenarios.

When combining the results of the POS-tagging with the Nosketch-Engine it is easy to browse corpora and even reuse POS-patterns for similar languages. This is especially useful for generic patterns for which only little knowledge about the specific language is required. This

includes knowledge about typical articles and conjunctions, or knowledge about word order as included in the World Atlas of Language Structures (Dryer & Haspelmath, 2013).

Figure 4 contains the output for such a query that identifies potential hyponyms using an English text corpus (Query: NOUN *"like"* NOUN *"and"* NOUN) .

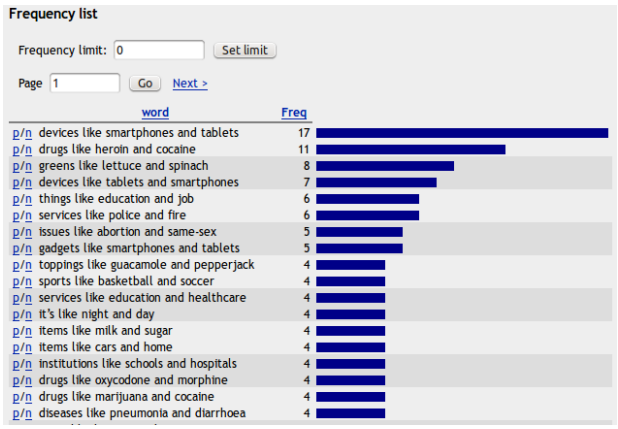

Figure 4: Sample of a generic corpus query using the NoSketch-Engine with UD17-tags for an English newspaper corpus.

3. Web Services for Corpus Querying

The emerging of integrated research infrastructures (like CLARIN[5] or DARIAH[6] in the European context) has given work on simplifying access to linguistic resources a boost. One major advantage of these infrastructures is the combination of existing data and tools to new workflows and applications with a manageable amount of effort. A precondition for these is the availability of and access to resources in a standardized matter via standardized access protocols. In the context of LCC a comprehensive set of RESTful Web services was deployed to give users access to the data[7]. Table 4 gives a shortened overview of available services.

To minimize the entrance barrier even further an OpenAPI-based specification[8] is provided that allows executing requests directly in a Web browser without the need of any programming experiences. Figure 5 shows a screenshot of this documentation for one service.

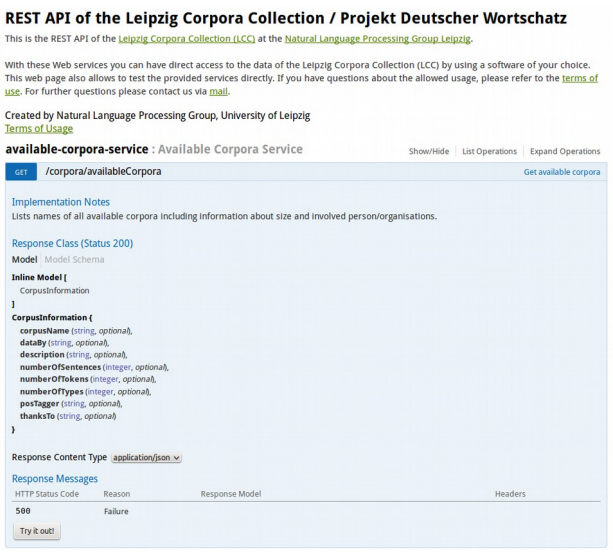

Figure 5: Documentation of the new RESTful API of the LCC using Open API/Swagger.

Name of Service	*Description*	*Availability*
LikeWords	Words corresponding to a pattern	All
Sentences	Reference sentences	All
Cooccurrences Sentences	Significant sentence cooccurrences for a word	All
WordAttributes	Various properties of words like hyphenation, affiliation to a subject area, description of the term etc.	Subset
WordRelations	Semantic relations between words in corpus (being part of the same synset, synonyms, antonyms etc.)	Subset
SimilarWords-Levenshtein	Words with Levenshtein distance	All
WordSets	Affiliation to a synset. (based on (Dornseiff, 2004))	German

Table 4: Overview of available Web services of the LCC (shortend).

5 https://clarin.eu
6 https://dariah.eu
7 http://wortschatzwebservices.informatik.uni-leipzig.de
8 https://openapis.org

4. Bibliographical References

Dornseiff, F. (2004). Der deutsche Wortschatz nach Sachgruppen. 8., völlig neu bearb. u. mit einem vollständigen alphabetischen Zugriffsregister versehene Aufl. von Uwe Quasthoff. Mit einer lexikographisch-

historischen Einführung und einer Bibliographie von Herbert Ernst Wiegand. Berlin. New York.

Dryer, M. S., Haspelmath, M. (eds.) (2013). The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2016-02-09.)

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1):61–74.

Heyer, G.; Läuter, M.; Quasthoff, U; Wittig, Th.; Wolff, Chr. (2001). Learning Relations using Collocations. In: A. Maedche, S. Staab, C. Nedellec and E. Hovy, (eds.). Proc. IJCAI Workshop on Ontology Learning, Seattle/ WA, 19. - 24. August 2001.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008) (pp. 425-432).

Kilgarriff, A., et al. (2014). The Sketch Engine: ten years on. In Lexicography: 1–30.

Petrov, S., Dipanjan D., McDonald, R. (2011). A universal part-of-speech tagset.arXiv preprint arXiv:1104.2086 (2011).

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing.Brno : Masaryk University, 2007. p. 65-70. ISBN 978-80-210-4471-5.