Finding Definitions in Large Corpora with Sketch Engine

Vojtěch Kovář[1,2], **Monika Močiariková**[1], **Pavel Rychlý**[1,2]
[1]NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic
[2]Lexical Computing CZ, Brno, Czech Republic
xkovar3@fi.muni.cz, 374485@mail.muni.cz, pary@fi.muni.cz

Abstract

The paper describes automatic definition finding implemented within the leading corpus query and management tool, Sketch Engine. The implementation exploits complex pattern-matching queries in the corpus query language (CQL) and the indexing mechanism of word sketches for finding and storing definition candidates throughout the corpus. The approach is evaluated for Czech and English corpora, showing that the results are usable in practice: precision of the tool ranges between 30 and 75 percent (depending on the major corpus text types) and we were able to extract nearly 2 million definition candidates from an English corpus with 1.4 billion words. The feature is embedded into the interface as a concordance filter, so that users can search for definitions of any query to the corpus, including very specific multi-word queries. The results also indicate that ordinary texts (unlike explanatory texts) contain rather low number of definitions, which is perhaps the most important problem with automatic definition finding in general.

Keywords: Sketch Engine, definition, definitions, CQL, corpora

1. Introduction

Definitions, in the sense of text descriptions of particular concepts, are an important part of any learning material, and especially dictionaries. At present, definitions in such materials are always hand-made, created by specialists in the particular field, or professional lexicographers.

This situation has two significant drawbacks – first, thinking up and writing the definitions is quite expensive; a specialist, or an educated lexicographer is needed for this work, and it is also quite an abstract and time-consuming task. Second, the knowledge of the particular specialist, and/or the time for thinking about the definition, may be limited and therefore the created definitions may not be good enough.

This paper presents a partial solution to these problems. From big data, in form of huge text corpora, we automatically extract sentences that may contain definitions and present them to a user (lexicographer/specialist) working on a particular concept; then the user has a possibility to simply re-use a particular definition, or adjust one of them according to others. The process of thinking about definitions is then made faster and more straightforward.

The work presented here is a rule based pattern matching approach developed and integrated within the Sketch Engine corpus query system (Kilgarriff et al., 2014a), primarily dedicated to lexicographers, so we are aiming mainly at lexicographic type of definitions. The whole exercise fits into a larger frame of streamlining production of dictionaries; after automatic terminology extraction (Kilgarriff et al., 2014b) finding typical collocations (Kilgarriff et al., 2004) and educational examples (Rychlý et al., 2008), definition creation is one of the last parts of the process without significant computational support.

In the paper, we briefly mention related work done on this task, then describe our implementation within the Sketch Engine, and provide a precision evaluation of this approach, for English and Czech (as an example of free-word-order language with rich morphology).

2. Related Work

Google definition boxes[1] (that appear sometimes when you search on Google) are very well-known result of "automatic" definition finding. However, they are probably (to the best of our knowledge, the exact algorithm has not been published yet) based on a few reliable sources of definitions, such as first paragraphs of Wikipedia articles or particular dictionaries. Definitions of concepts not covered by these resources, but present on the web, cannot be found easily using Google. Also, only one definition candidate, or one dictionary record is provided regardless the context wanted by the user.

We do think that our approach in combination with large web corpora (Jakubíček et al., 2013) or specialized corpora (e.g. created from web by the WebBootCat tool (Baroni et al., 2006)) will perform better in this regard.

Other related work includes academic systems for finding definitions (Klavans and Muresan, 2001; Navigli et al., 2010; Jin et al., 2013), using manually created rules, machine learning or a combination of both. The results are quite satisfying – e.g. (Jin et al., 2013) reports 92% precision and 79% recall.

For Slavic languages, there is a report (Przepiórkowski et al., 2007) indicating that the situation is much worse than in English – the evaluation sets are rather small and the resulting figures significantly lower – around 20% in precision and 40% recall.

We are not trying to compete with these systems and numbers – our aim was to build the definition finding functionality into the Sketch Engine, so that it is fast enough to be used with very huge corpora, technically compatible with the rest of the system and highly customizable. The definition finding function should extract as many correct definitions as possible while keeping reasonable precision. Such parameters should offer a decent number of definition candidates for each term (provided the definitions are there in the corpus) where most of them are correct.

[1] `www.googleguide.com/dictionary.html`

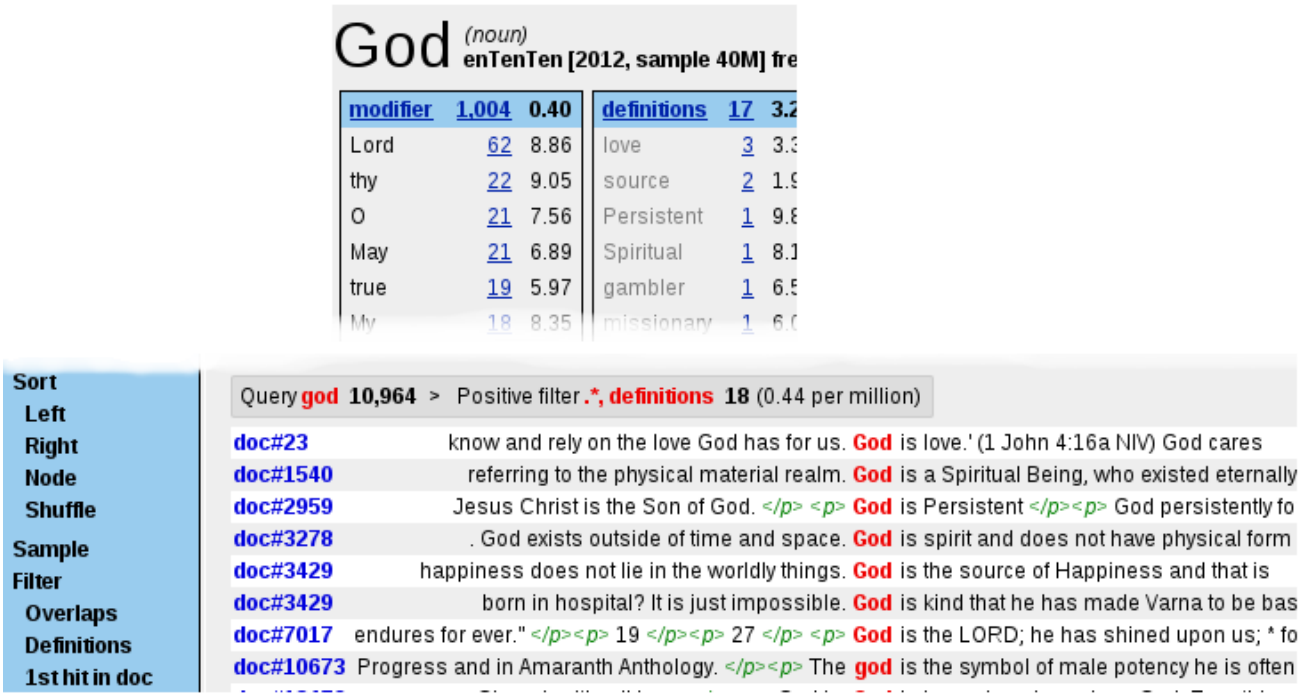

Figure 1: Definitions among other word sketch relations, and an example of definition candidates for a particular word ("God") – link to definitions is under "Filter" near the bottom of the concordance menu.

```
( (<s> | (<s>[tag="DT" & lc!="this|these|those"]) |
   ([tag!="IN|PP.*|POS" | lc="while|although"]
             [tag="DT" & lc!="this|these|those"]) |
   ([tag!="DT|IN|PP.*|POS|N.*|JJ.*|VVG|CD" | lc="while|although"])
  )
  ([tag="N.*|JJ|VVG|CD"]{0,3} !containing
   (meet [tag="N.*|JJ|VVG"] [tag="IN" & lc!="while|although"] -1 0)
  )
  1:[tag="N.*" & lemma!="reference|use|...|name|definition"]
  "\"|'"?
  "is|are"
  [tag="RB"]?
  "understood"
  "to"
  [tag="VV|VB"]
  [tag!="N.*"]{0,12}
  2:[tag="N.*"])
) within  <s/>
```

Figure 2: Example definition pattern in CQL, for **TERM "is/are" "understood" "to"**.

3. Implementation

We have based the definition extraction algorithm on the word sketch formalism (Kilgarriff et al., 2004) which exploits the queries in the corpus query language (CQL)[2] to find patterns in the corpus. In case of finding collocations, the patterns recognize pairs of words and the results are sorted according to frequency or an association score (Rychlý, 2008).

We implemented definition finding as another word sketch relation – that is, we described the most common definition patterns using CQL, and indexed them together with word sketch information, so that it is easily accessible even for billion-size corpora. The patterns were inspired by many sources – apart from the ones cited earlier, personal correspondence with Michael Rundell, one of the world's leading experts in lexicography, was especially valuable.

Then we connected the indexes with concordance search, as illustrated in Figure 1. A new link called "Definitions" is a pre-defined filter of any concordance result, that intersects the particular results with the word sketch indexes

[2] www.sketchengine.co.uk/corpus-querying

Pattern type	No. hits	Prec. on sample (%)	Estimated no. definitions
is/are/...	1,751,813	77	1,342,215
what is	1,574	16	251
refers to	40,690	57	23,093
is defined as	54,435	29	15,781
is known as	78,756	54	42,528
is used to describe	3,934	63	2,485
is a term for	11,504	74	8,512
is understood to	618	35	216
consists of	12,821	29	3,721
Total	1,956,145	74	1,438,802

Table 1: Precision and estimated number of correct definitions found, according to pattern groups. English Wikipedia.

Corpus	No. hits	Prec. on sample (%)	Estimated no. definitions
enTenTen (40M)	21,833	57	12,379
Czech Wikipedia	105,210	73	76,632
czTenTen	237,890	31	73,746

Table 2: Overall precision and estimated number of correct definitions found for the 3 other corpora.

for definitions. As the result, only the definition candidates matching the current query are displayed.

3.1. Patterns

The CQL patterns used for matching definition candidates can be described in a simplified way, and summarized, as follows:

- **TERM "is/are/means/was/were" "a/an"**, including:
  - **"TERM"** (in quotes)
  - **TERM parenthesis "is/are/..." "a/an"** (parenthesis expressed by commas, dashes or brackets)
  - **TERM prepositional-phrase "is/are/..." "a/an"**
  - **optional "a/an"** in selected cases
- **"What" "is" TERM**, with a definition in the following sentence
- **TERM "refers" "to"**, plus variants with parentheses and prepositional phrases, as above
- **TERM "is/are" "defined" "as"**, plus variants with parentheses and prepositional phrases, as above
- **... "is/are" "known/called/referred_to" "as" TERM**
- **TERM "is/are" "used" "to" "describe/denote/mean/refer_to"**, plus variants with parentheses
- **TERM "is" "a" "term" "for/referring_to"**, plus variants with parentheses
- **TERM "is/are" "understood" "to"**, plus variants with parentheses
- **TERM "consists" "of"**, plus variants with parentheses

In all cases, **TERM** stands for a general noun phrase. The particular CQL queries are rather complex in most cases (but they can be made more readable by using macros), and exclude some particular expressions that indicate non-definitions. An example for **TERM "is/are" "understood" "to"** is shown in Figure 2.

The patterns for the Czech language are mostly similar, exploiting the translations of the expressions above, and taking into account the syntax of the language. Similar "translations" would be probably doable for other languages.

In total, there were 50 CQL patterns developed for English and 37 for Czech.

4. Evaluation

The patterns were developed on part of the English Wikipedia corpus available in the Sketch Engine (1.4 billion words), then tested on the full English Wikipedia corpus (using random samples, see below)[3] and then on a 40 million sample of the enTenTen corpus.[4] In case of Czech, we used the Czech Wikipedia corpus (60 million words) and the czTenTen web corpus (5 billion words).

For each pattern, a random sample of 50 hits was selected, and percentage of correct definitions was counted by an annotator. This percentage was then extrapolated to the whole corpus and number of correct definitions found by the pattern was estimated. Results for the English Wikipedia corpus are summarized in Table 1, according to pattern groups described above.

Overall results for the other corpora are summarized in Table 2. For the TenTen corpora, the same method was used but all the patterns were evaluated at once using a bigger sample of 200 hits.

[3] In some cases, there may be an intersection between development and evaluation sets, but due to the number of results for the whole corpus – about 2 million for all the patterns – this intersection is absolutely marginal.

[4] In this case, there was no possible intersection.

38 of the 50 English patterns reached precision over 50 %, for Czech it was 33 of 37.

4.1. Discussion

We can see that the distribution among the patterns is very uneven, over 90% of the estimated correct definitions were matched by the most common pattern group. Also, the procedure yields much more results on Wikipedias than general internet texts, and also precision varies. This is probably caused by the fact that general texts are quite poor on definitions.

The Czech and English Wikipedias yield proportionally similar numbers of correct definitions, also the precisions are nearly the same; this indicates that the two pattern sets for two different languages are of similar quality. On the other hand, the Czech web corpus results are much worse than in English – we do not have a good explanation of this interesting observation, perhaps the Czech internet contains less educative texts.

5. Conclusion

We have introduced a method for finding definitions in large text corpora, using a pattern matching approach. The procedure exploits features of the Sketch Engine corpus query system, its query language, indexing and filtering options.

The results indicate that the method is able to extract a large number of correct definitions from general language corpora, with reasonable precision to be useful for lexicographers and other definition finders.

Acknowledgements

This work has been partly supported by the Grant Agency of CR within the project 15-13277S. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047. Last but not least, we would like to thank Michael Rundell for his priceless consultancy on the topic of the paper.

6. Bibliographical References

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCat: a web tool for instant corpora. In *Proceeding of the EuraLex Conference 2006*, pages 123–132, Italy. Edizioni dell'Orso s.r.l.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference CL 2013*, pages 125–127, Lancaster. Lancaster University.

Jin, Y., Kan, M.-Y., Ng, J.-P., and He, X. (2013). Mining scientific terms and their definitions: A study of the ACL anthology. *EMNLP-2013*.

Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. *Information Technology*, 105:116–127.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014a). The Sketch Engine: ten years on. *Lexicography*, 1.

Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P., and Suchomel, V. (2014b). Finding terms in corpora for many languages with the Sketch Engine. In *Proceedings of the Demonstrations at the 14th Conferencethe European Chapter of the Association for Computational Linguistics*, pages 53–56, Gothenburg, Sweden. The Association for Computational Linguistics.

Klavans, J. L. and Muresan, S. (2001). Evaluation of the DEFINDER system for fully automatic glossary construction. In *Proceedings of the AMIA Symposium*, page 324. American Medical Informatics Association.

Navigli, R., Velardi, P., and Ruiz-Martínez, J. M. (2010). An annotated dataset for extracting definitions and hypernyms from the web. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Przepiórkowski, A., Degórski, Ł., Wójtowicz, B., Spousta, M., Kuboň, V., Simov, K., Osenova, P., and Lemnitzer, L. (2007). Towards the automatic extraction of definitions in Slavic. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 43–50. Association for Computational Linguistics.

Rychlý, P., Husák, M., Kilgarriff, A., Rundell, M., and McAdam, K. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*, pages 425–432, Barcelona. Institut Universitari de Lingüística Aplicada.

Rychlý, P. (2008). A lexicographer-friendly association score. In *Proceedings of Second Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 6–9, Brno. Masaryk University.