Practical Post-Editing Lexicography with Lexonomy and Sketch Engine

Milos Jakubicek, Michal Měchura, Vojtech Kovar, Pavel Rychly
Lexical Computing, Natural Language Processing Centre, Masaryk University, Faculty of Informatics, Masaryk University
milos.jakubicek@sketchengine.co.uk, valselob@gmail.com, xkovar3@fi.muni.cz, pary@fi.muni.cz

In this paper we present an implemented solution for post-editing dictionaries within the Lexonomy dictionary writing system interconnected with the Sketch Engine corpus management system. We follow up on the work on automatic dictionary drafting ("One-Click Dictionary") and assume a dictionary draft to be post-edited, or an existing (edited) dictionary to be extended are in place. We cover focus on features that enable lexicographers to obtain relevant corpus evidence or corpus-driven analysis in a user-friendly way from within the environment of the dictionary writing system and thus speed up their workflow. We exemplify the usage scenario in an ongoing project on Danish-English-Korean bilingual dictionary.

Introduction

This paper focuses on ongoing trends in automatizing the lexicographic processes with the help of advanced natural language processing techniques applied on top of large text corpora. By now corpora are at hands of lexicographers for over two decades, representing sources of empirical evidence. The influences were mutually beneficial – lexicography has significantly contributed to corpus development as well as to advances in corpus linguistics. While at the beginning corpora were primarily used for manual inspection of data by lexicographers, many advances in natural language processing allow for extended automation of the lexicographic process where lexical information is first automatically obtained from corpora and later post-edited by lexicographers.

We first briefly describe the two systems where the presented techniques are implemented – Sketch Engine corpus management system and Lexonomy dictionary writing system – and then show how they are used for creating initial dictionary drafts automatically and/or post-editing of existing dictionaries and dictionary drafts.

Sketch Engine

Sketch Engine (Kilgarriff et al., 2014) is a leading corpus management system primarily designed for lexicographic purposes. It is web-based and provides access for several hundreds of corpora for (as of December 2017) over 90 languages. It has been used for lexicographic projects for dozens of languages[1] and implements several analytic functions useful for dictionary drafting. These include single- and multiword retrieval, automatic extraction of good dictionary examples (GDEX), automatic extraction of definitions (GDEF) or collocational analysis of words' behaviour known as "word sketches" which serves as a backbone for obtaining a distributional thesaurus or word sense clustering.

1 For an overview please see https://www.sketchengine.co.uk/bibliography-of-sketch-engine

Sketch Engine has a separate corpus building part which allows users to build corpora from their own texts, or from texts automatically retrieved from the web according to their specifications taking the form of seed words which are passed to a search engine to find relevant online results (the so called WebBootcat approach). User corpora can be subject to automatic terminology extraction in order to retrieve domain-specific lexical items suitable for headword list generation.

Lexonomy

Lexonomy (Měchura, 2017) is a simple web-based dictionary writing system integrating a dictionary publishing platform. It allows its users to devise custom dictionary structures (XML templates) using a simple graphical editor and subsequently import entries or manually create and edit new ones. Lexonomy is interconnected with Sketch Engine using two interaction modes: the push model and the pull model.

In the push model, Sketch Engine pushes an initial dictionary draft fully automatically into a new dictionary in Lexonomy using its API access. In the pull model, users of Lexonomy retrieve different entry parts such as corpus examples from Sketch Engine (again, using its API access) without leaving the Lexonomy interface.

One-Click Dictionary

The push model was exploited to implement a One-Click Dictionary feature in Sketch Engine (see Jakubíček et al., 2017) which exports a dictionary draft based on a corpus selected by the user. Depending on the annotation available for the respective corpus, the export entails following features:

- headword list generation
- collocation extraction
- word sense clustering
- example sentences
- definitions
- thesaurus

While some of these functions (e.g. collocation extraction) achieve very high accuracy, the entry still needs to be manually post-edited. In fact, efficient implementation of the post-editing is key to the success of this approach: post-editing must never be more time consuming for the lexicographer than writing the entry from scratch.

Post-Editing in Lexonomy

The post-editing features in Lexonomy facilitate the pull model for interacting with Sketch Engine. They are usable for dictionaries that were initially drafted automatically from Sketch Engine as well as any other ones present in Lexonomy, regardless whether they were imported or manually created. They include the following operations:

- evaluation of lemma coverage against a corpus and retrieval of additional headwords
- retrieval of examples sentences using GDEX
- retrieval of definitions using GDEF

- retrieval of collocations
- word sense clustering
- retrieval of thesaurus items

Each of these steps contains an interface where editors can easily accept or reject the data as retrieved from Sketch Engine; or post-edit it (including lumping and splitting of word senses).

Use case: Danish-English-Korean

The post-editing workflow as described is at the moment exploited in a commercial project for creating a Danish to English and Korean bilingual dictionary. In the final paper we will provide feedback from editors and report on overall success of this approach within the project.

Conclusions

In this paper we show a practical implementation of the post-editing approach based on Sketch Engine and Lexonomy. The main goal of this strategy is to foster corpus use and ease the access to corpus evidence while advancing the automation of the dictionary creation process and speed up editors' work. We discuss the post-editing workflow of particular entry parts as well as experience from an existing project on Danish-English-Korean bilingual dictionary.

Keywords: corpora, post-editing, Lexonomy, Sketch Engine