Linguistic Linked Open Data as a Source for Terminology - Quantity versus Quality

Shanshan Wang

Lauri Carlson

University of Helsinki

This paper studies the accessibility and quality of linguistic linked open data (LLOD) for multilingual terminology and localization. Terminologists may benefit from LLOD due to its openness, freedom of copyrights, and multilingual and multi-domain coverage. The paper mainly discusses the quality of LLOD data for terminological use, which we are in the process of studying with our ontology-based terminology tool TermFactory, designed to search, merge, verify and improve the quality of linked term data.

Primary sources for linked open data (LOD) include open collaborative platforms like Wikipedia and Wiktionary, and data converted from projects put in common domain like WordNet. RDF datastore conversions of the primary sources, such as DBPedia, can be used for multilingual terminology management, localization and translation. Since the data are collected from open and non-commercial sources, the data suffers from unreliability, uneven coverage, redundancy and ambiguity.

In the paper, we study a large derivative multilingual lexical data sources: BabelNet 3.0 to analyse and evaluate its usefulness for professional special language terminology. BabelNet 3.0 contains more than 13 million entries called Babel synsets, which represent a given meaning and contains all the synonyms which express that meaning in a professed range of 271 languages. It is both a multilingual encyclopedic dictionary and a semantic network. We take 9 test English terms in 3 domains: medical, law and social network including both general languages terms and specific language terms. The aim is to test the coverage and reliability of BabelNet data and compare with other sources such as IATE, Eurotermbank, Wiktionary and Termwiki.

For this note, we trawled the terms from 5 multilingual lexical data sources and counted entries (concepts, synsets, meanings) per key, the number of terms (senses) per entry, and the number equivalents (synonyms, translations) per entry. The query and basic statistics done, we are in the process of comparing ways to match senses between the different sources. Later on, we plan to extend our data and run or methods on a more representative sample.

As for the statistical query results, most of the test terms could be queried in BabelNet due to its large quantity of terms and domain coverage, and multilingual coverage. It is a well constructed LOD and machine translation tool. The existing problems include uneven quality especially in minor language translations, errors and duplicates in common language words.

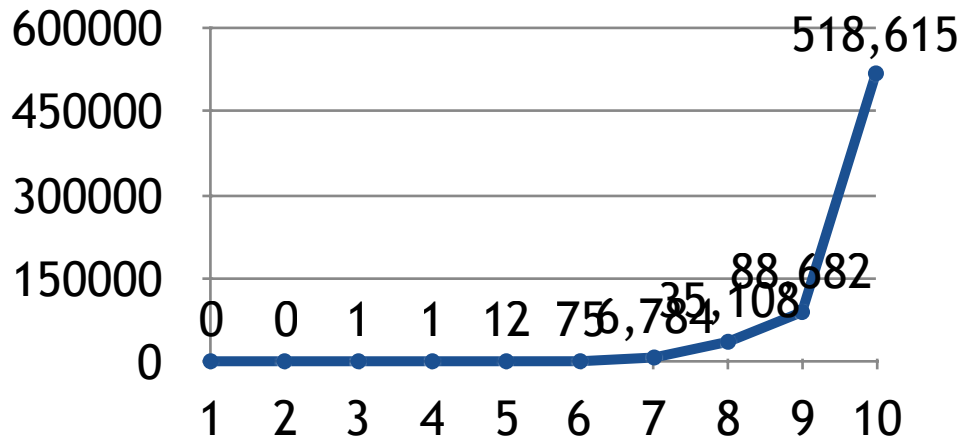
We chose our mini sample with reference to token frequencies from two large corpora, GloWbE (Corpus of Global Web based English) and Google book. We select the following three terms from three fields each: medicine, law, and web: medicine: *aphagia, colporrhaphy, tympanoplasty*; law:*implied in-fact contract, indictment, nuncupative will*; web: *blogger, feed, post*. The medicine terms are monosemic latin/greek coinages. The law terms are collocations containing some special lexemes.

One of the web terms is monosemic, two are common language words with field specific meaning.

Terms are needles in a haystack in two ways. For narrow specialty terms, the problem is to find the term at all, and if found, to know if the source can be trusted. For this type, term banks still beat corpora. For terms narrowed from common language, the problem is to identify the specialist sense. For this end, the more information there is about the entry in the source, the better. With this in mind, we also produced statistics on the number of features (converted to RDF properties) each source provides.

GloWbE frequencies:

1:implied in-fact contract 2:aphagia 3:nuncupative will 4:colporrhaphy 5:memorandum of law 6: tympanoplasty 7: indictment 8: blogger 9:feed 10:post



Google book frequencies:

2006	
memorandum of law (All)	0.0000031808%
indictment (All)	0.0003728161%
nuncupative will (All)	0.0000007216%
aphagia (All)	0.0000004180%
colporrhaphy (All)	0.0000050880%
tympanoplasty (All)	0.0000066229%
blogger (All)	0.0000540348%
feed (All)	0.0026545580%
post (All)	**0.0108030071%**

We trawled the sources for the sample keys (specifying the string and the language code) using our TermFactory toolkit. The kit downloads for each website the first page retrieved for the key plus any detail pages directly accessible from it, tidies the HTML pages into XML and runs a XSLT converter to transform the data into RDF. (We use the same method for Babelnet for fairness' sake. Otherwise, we would prefer querying the Babelnet SPARQL service.) Here are statistics for the total number of triples, number of languages, and number of distinct RDF properties caught in the trawl. (The results are tentative in that we cannot promise that each source got fully exhausted.

Term bank	Triples	Language code	Property
BabelNet	15203	12	18
ETB	472	30	4
IATE	3265	25	11
Termwiki	280	1	5
Wiktionary	1775	70	25

BabelNet and other LOD in general do well in the coverage and multilingual translations for the special terms, while traditional terminology databases still hold the lead in quality, reliability and provenance information.

	BabelNet:id	ETB	IATE:id	termwiki:id	wiktionary:gloss
aphagia@en	1	0	1	1	1
tympanoplasty@en	2	1	1	1	0
colporrhaphy@en	1	2	2	0	0
memorandum of law@en	1	0	1	0	0
numcupative will@en	1	0	1	1	0

	BabelNet:id	ETB	IATE:id	termwiki:id	wiktionary:gloss
indictment@en	2	11	6	9	3
implied-in-fact contract@en	0	0	0	0	0
blogger@en	2	6	1	3	1
post@en	22	40	7	25	17
feed@en	14	32	10	16	8

Babelnet

entries are WordNet like synsets identified by babelnet id (synsetId). senses are identified by property babelnet:sense.
Babelnet applies a Linked data approach. Entries are linked partly automatically, which may cause uneven quality, errors and duplicates. Provenance and reliability information is not always available.

eurotermbank

entries are identified by source (collection) and field (subject). EuroTermBank data apparently largely comes from IATE. Exhaustive queries from ETB were slow. Really rare terms in our sample were not found. The common terms have many duplicates. . EuroTermbank is a federated source, which may explain some of the above observations.

termwiki

entries are terms identified by termwiki:id (iid).
termwiki is a commercial crowdsourced glossary-type term site. As such, the quality varies. TermWiki is multilingual, though English predominates. *) Translations were not yet trawled.

wiktionary

entries are word senses identified by etymon, part of speech and gloss.
Wiktionary is a crowdsourced open-source general language dictionary. As such, does not provide provenance and reliability information. The most multilingual of our sample, contains grammatical information (not trawled yet).

References

C. Chiarcos, S. Hellmann, and S. Nordhoff. 2012. Linking lin- guistic resources: Examples from the open linguistics working group. In C. Chiarcos, S. Nordhoff, and S. Hellmann, editors, Linked Data in Linguistics. Representing Language Data and Metadata, pages 201–216. Springer

C. Chiarcos, J. McCrae, P. Cimiano, and C. Fellbaum. 2013. Towards open data for linguistics: Linguistic linked data. In A. Oltramari, P. Vossen, L. Qin, and E. Hovy, editors, New Trends of Research in Ontologies and Lexical Resources, pages 7–25. Springer

Cimiano, Philipp, et al. "Linked terminology: applying linked data principles to terminological resources." *Proceedings of eLex* (2015).

Fontenelle, Thierry, and Dieter Rummel. "Term banks." *International Handbook of Modern Lexis and Lexicography*. Springer Berlin Heidelberg, 2014. 1-12.

Henriksen, Lina, Claus Povlsen, and Andrejs Vasiljevs. "EuroTermBank–a Terminology Resource based on Best Practice." *Proceedings of LREC 2006, the 5th International Conference on Language Resources and Evaluation*. 2005.

J. McCrae, E. Montiel-Ponsoda, and P. Cimiano. 2012c. Collab- orative semantic editing of linked data lexica. In Proceedings of the 8th International Conference on Language Resource and Evaluation, pages 2619–2625, Istanbul, Turkey.

Lušicky, Vesna, and Tanja Wissik. "Terminology: Don't only collect it, use it!." *Aslib Translating and the Computer Conference*. Vol. 34. 2012.

Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John McCrae, Philipp Cimiano, Roberto Navigli, "Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0." Proc. of the 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, 26-31 May, 2014.

S. Auer, J. Lehmann, and A. N. Ngomo. 2011. Introduction to linked data and its lifecycle on the web. In Reasoning Web. Se- mantic Technologies for the Web of Data, pages 1–75. Springer.