Powering Complex Business Signals' Classification using Enriched Taxonomy by Existing Data Sources

Muhammad Arslan, Christophe Cruz

Powering Complex Business Signals' Classification using Enriched Taxonomy by Existing Data Sources

M. Arslan[1] and C. Cruz[2]. (1,2) Univ. Bourgogne Franche-Comté, LIB, EA 7534, 9 Avenue Alain Savary, Dijon, France, [*]muhammad.arslan@u-bourgogne.fr*

Taxonomies play an important role in Business Intelligence (BI) applications by capturing business signals to facilitate decision support processes [1]. The business signals are complex as they are broad-ranging related to office expansions or relocations, executive changes or major hires, partnerships, mergers, acquisitions, etc. carrying diverse information related to future transcendental changes [1, 2]. Capturing and classifying these changes as business signals using online news sources for different interested parties is crucial for making company decisions at the right time. Conventionally, the classification of news documents is performed using detected keywords of the taxonomy. A taxonomy is a tree-based structure where nodes (concepts) are linked with the hypernymy relation [2]. BI processes require knowledge of the fundamental business concepts (i.e. defined as abstract ideas) organized as taxonomies for effective decision-making. To ensure and maintain reliable data quality, it is crucial to keep the same definitions and organization of these concepts. This highlights the major significance of business taxonomies in the BI domain. However, their development in business information systems follows an ad hoc process in the majority of the cases. Existing studies do cover BI taxonomies but these are excessively generic and domain-specific [1 - 3]. As a result, the BI domain suffers from many immature, incorrect, and incomplete notions of concepts. The contribution of our research is the presentation of a method (see Figure 1) based on existing data sources to enrich the human-made taxonomy with new concepts to improve business news classification. To enrich a hand-crafted taxonomy of a company consisting of 2,973 concepts, 3 datasets are used which are; 1) WordNet [4], 2) Sense2vec [5], and 3) Wiktionary [6] (mentioned in Table 1).

Table 1: Statistics of datasets used in this study.

No.	Dataset	Nouns	Adjective	Verb
1	WordNet 3.0 [4]	117,798	21,479	11,529
2	Sense2vec (file: *s2v_reddit_2015_md*) [5]	816,550	34,365	66,599
3	Wiktionary [6]	218,629	62,202	58,872

WordNet is an online lexical database that provides a combination of traditional lexicographic information. English nouns, verbs, adjectives, and adverbs are organized into sets of synonyms, each representing a lexicalized concept. However, WordNet mostly contains data on 1-word concepts. To deal with taxonomy concepts having 2 to 3 words (e.g. Public relations, Financial advice, etc.), Sense2vec model is used. It is a model for word sense disambiguation based on supervised Natural Language Processing (NLP) labeling that is used to disambiguate between word senses. Lastly, for the words which were not found in the WordNet, 2) Sense2vec datasets,

their data is extracted from the Wiktionary. Using these datasets, around 58% of taxonomy concepts are enriched (see Table 2).

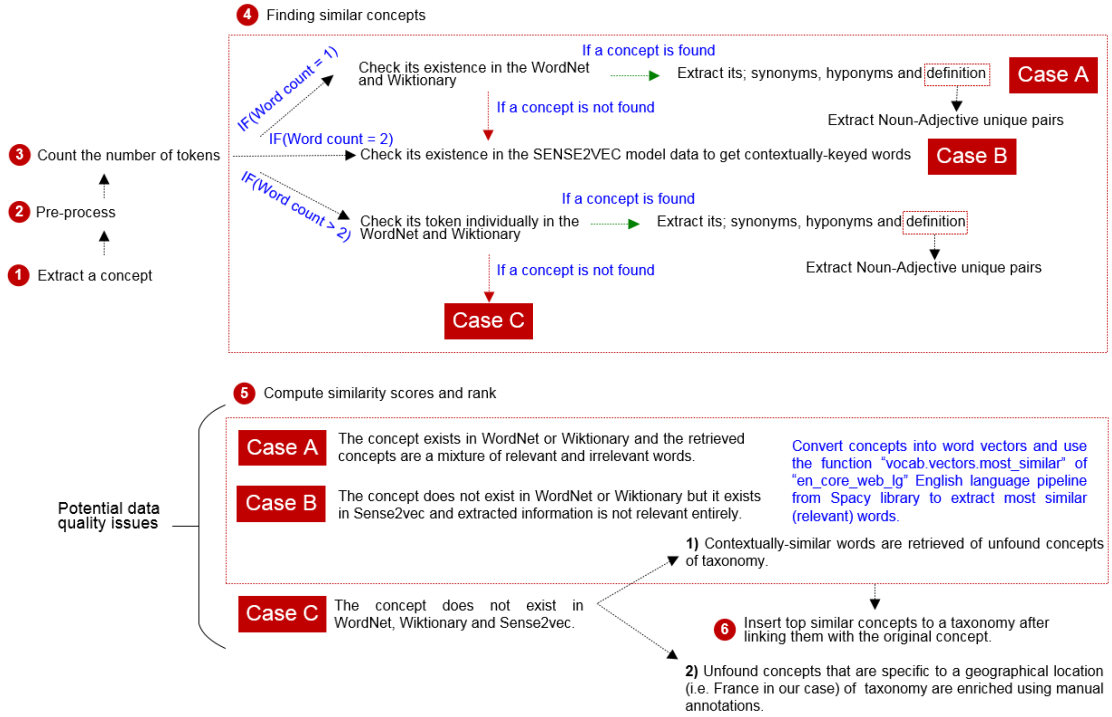

Figure 1: Taxonomy enrichment method

The benefit achieved from the above-mentioned taxonomy enrichment approach is that this approach has significantly reduced the computational overhead of word-sense modeling. It provides a simple mechanism for taxonomy enrichment tasks to select the appropriate sense embedding without training the models to find similar words. Consequently, enriched taxonomy will help to improve the classification of news documents.

Table 2: Achieved taxonomy enrichment.

Total taxonomy concepts	Enriched	Not found	New concepts added	Enrichment achieved (%)
2,973	1,729	1,244	40,212	58

Acknowledgements

The authors thank the French company FirstECO (https://www.firsteco.fr/) for providing the taxonomy and the French government for the plan France Relance funding.

References

[1] H. Angermann, TaxoMulti: Rule-based expert system to customize product taxonomies for multi-channel e-commerce, *SN Computer Science* 3 (2), (2022) 1-18.

[2] T. Schoormann, F. Möller, D. Szopinski, Exploring purposes of using taxonomies, *Wirtschaftsinformatik Proceedings* 5 (2022).

[3] I. Nikishina, et al., Taxonomy enrichment with text and graph vector representations, *arXiv preprint* (2022) 2201.08598.

[4] G.A. Miller, WordNet, An electronic lexical database, *MIT press* (1998).

[5] A. Trask, P. Michalak, J. Liu, sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings, *arXiv preprint,* (2015) 1511.06388.

[6] E. Navarro et al., Wiktionary and NLP: Improving synonymy networks, In *ACL Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources* (2019) 19-27.