language and translation technology team

WORD SENSE DISAMBIGUATION FOR SPECIFIC PURPOSES: AN EXAMPLE SENTENCE-BASED METHODOLOGY FOR INTELLIGENT COMPUTER-ASSISTED LANGUAGE LEARNING

JASPER DEGRAEUWE & PATRICK GOETHALS

ejercicio – Spanish WordNet – 5 synsets	*facturación* – Spanish WordNet – 1 synset
[...] 4 [ejercicio_anual – ejercicio] 5 [curso – ejercicio – año]	1 [programación – facturación]
Too fine-grained: clear-cut, useful difference between 12-month accounting period and regular 12-month period?	**Incomplete**: *facturación* as “check-in” (e.g. of luggage) and *facturación* as “revenue"?

1 STARTING POINT

- Corpora-based resources for language learning and teaching are usually non-disambiguated

→ Complicates generating vocabulary learning exercises (e.g. querying corpora for usage examples of specific semantic uses of vocabulary items)

2 CHALLENGES

- Choice of sense inventory tailored to the needs of the specific purpose
  - Large lexical/semantic resources available, both monolingual (e.g. WordNet[1]) and multilingual (e.g. BabelNet[2])
  - BUT: suitability of sense distinctions?
    - Too fine-grained and or incomplete for many NLP applications[3]
    - A single set of word senses is unlikely to be appropriate for different NLP applications: different corpora and different purposes lead to different senses[4]
  - Develop WSD methodology based on customised sense inventory
- Human-computer interaction in the classroom
  - Integration in learning module (make “intelligence” of ICALL accessible and understandable to users, and at the same obtain data to improve methodology)

3 WSD METHODOLOGY

- Specific purpose: ICALL for Spanish as a foreign language
- Experiment: 25-item lexical sample (business vocabulary)
- Elaboration of customised sense inventory
  - Divide input from Clave dictionary (SM Diccionarios) into main senses
  - Represent each sense by prototypical usage examples (available in Clave)
- Use pretrained BERT [5] model to convert example sentences into “sense embeddings” (= contextualised embeddings of ambiguous item)
- Calculate cosine similarity between sense embeddings and target sentences (= contextualized embedding of ambiguous item in those sentences)
- Classifier: select class of top match amongst cosine similarity values

4 RESULTS (F1, weighted average)

- Setup 1: only original example sentences from Clave as labelled input
- Setup 2: original example sentences as seed sentences to bootstrap labelled dataset
- Setup 3: original example sentences + **limited but targeted set of manually annotated example sentences** to bootstrap labelled dataset

Setup 1	Setup 2	Setup 3	MFS
0.84	0.86	0.89	0.57

***ejercicio* – 4 main senses**

1 realisation (of a right, profession, etc.) → *Era la dueña y entró en la finca haciendo* ***ejercicio*** *de sus derechos.*

2 physical exercise → *Si quieres perder peso, tendrás que hacer* ***ejercicio****.*

3 activity, task or practical test → *Intentar recordar algún poema es un buen* ***ejercicio*** *para la memoria.*

4 period of time, generally of 12 months → *Cerramos el último* ***ejercicio*** *con ganancias.*

GHENT UNIVERSITY

REFERENCES

[1] Fellbaum, C. (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

[2] Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217-250.

[3] Hovy, E., Navigli, R., & Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194, 2-27.

[4] Kilgarriff, A. (1997). I don’t believe in word senses. *Language Resources and Evaluation*, *31*(2), 91–113.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171-4186.