How's Business Going Worldwide ? A Multilingual Corpus for Business Relation Extraction

Institut de Recherche en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

Khaldi Hadjer [1,2], Benamara Farah [2], Pradel Camille [1], Sigel Grégoire [1], Aussenac-Gilles Nathalie [2]

[1]Geotrend, Toulouse, France [2]IRIT/CNRS, Toulouse, France

Motivation

The business world has changed due to the 21st century economy borders have melted and trades became free competition is no longer at the local market level but at the global level.

→ World Wide Web has become a major source of information for companies and professionals to keep track of their complex, rapidly changing, and competitive business environment.

Business Relation Extraction

Automating the extraction of multilingual economic and financial information while relaying on Information Extraction techniques, such as business relation extraction (BRE) from web content.

BRE aims at discovering either *Inner-Organizational* (Inner-ORG) relations linking a company and its components (e.g. company-employees, company-CEO) or *Inter-Organizational* (Inter-ORG) relations involving different companies (e.g. company-customer, company-partner)[7].

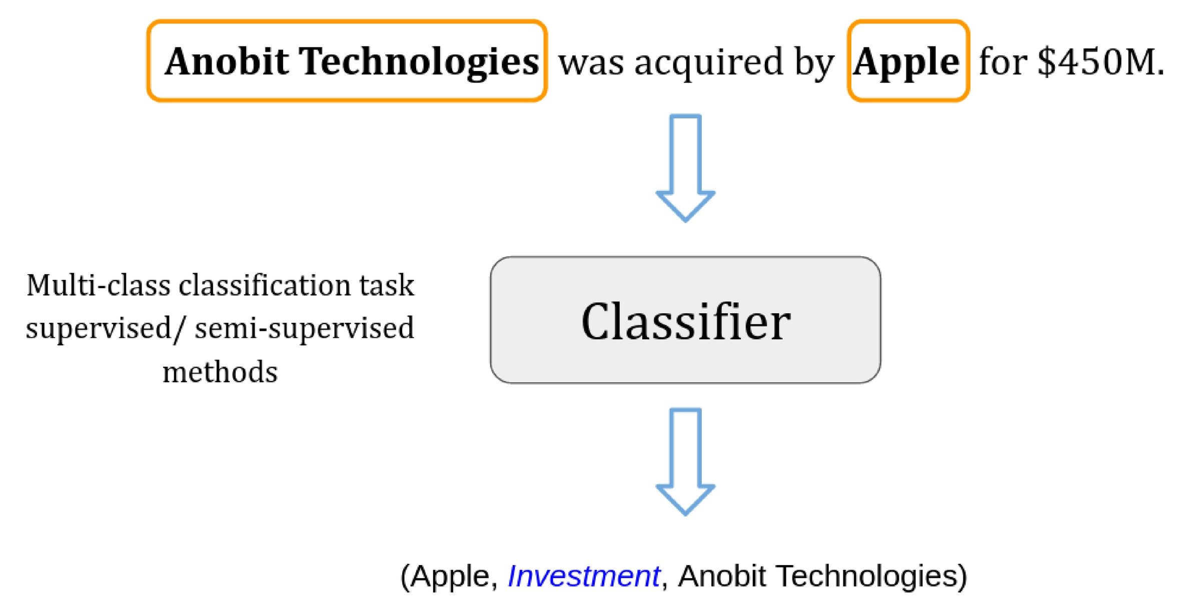

Figure 1. Business Relation Extraction.

Limitation of current works

Overall, BRE studies share three main limitations:

- Datasets used to train the models are either small or not freely available to the research community.
- Mainly, only two relations are considered, namely *Competition* and *Cooperation*.
- All the proposed models are monolingual targeting either [2, 7, 8], German [1], Chinese [6], or Portuguese [4, 3].

Main contributions

In this paper, we focus on **binary Inter-ORG BRE** from online web content, our contributions are as follows:

- **A unified characterization for Inter-ORG relations** focusing on five relations: Investment, Cooperation, Sale, Supply, Competition, and Legal proceedings.
- **BizRel, the first manually annotated multilingual dataset** annotated according to this characterization and considering four languages: *French*, *Spanish*, *English*, and *Chinese*.
- **A set of deep learning experiments** to detect these relations in multilingual texts.

Data Collection

We follow the procedure described in [5] for English business relations and extend it to French, Spanish, and Chinese. Data collection consists in extracting from the web relevant sentences following a three-step procedure:

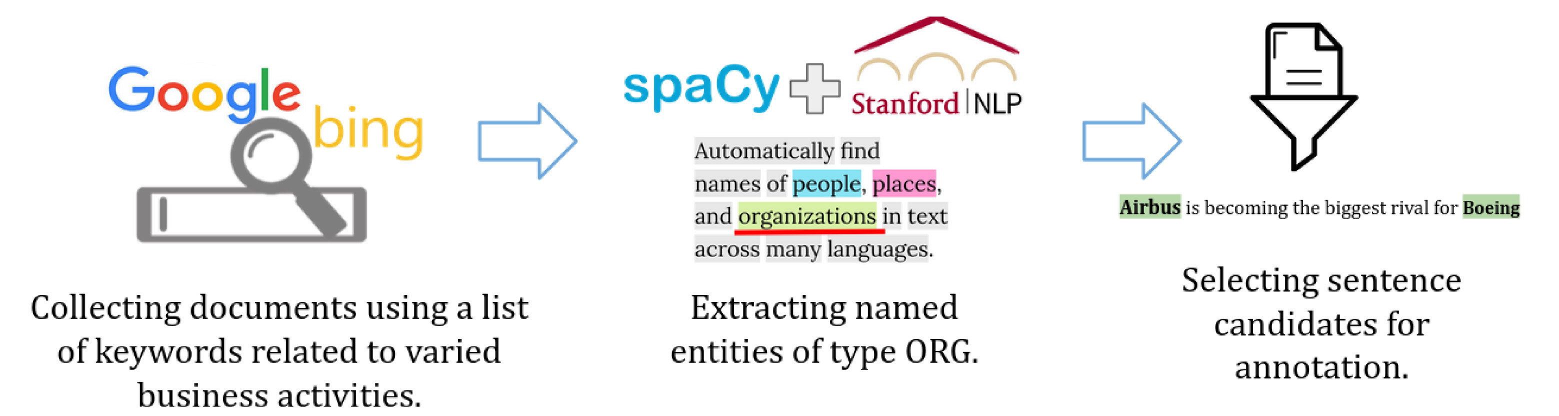

Figure 2. Data Collection Procedure.

This procedure resulted in a total of $25,469$ sentences for French, English, Spanish, and Chinese.

Characterizing Business Relations

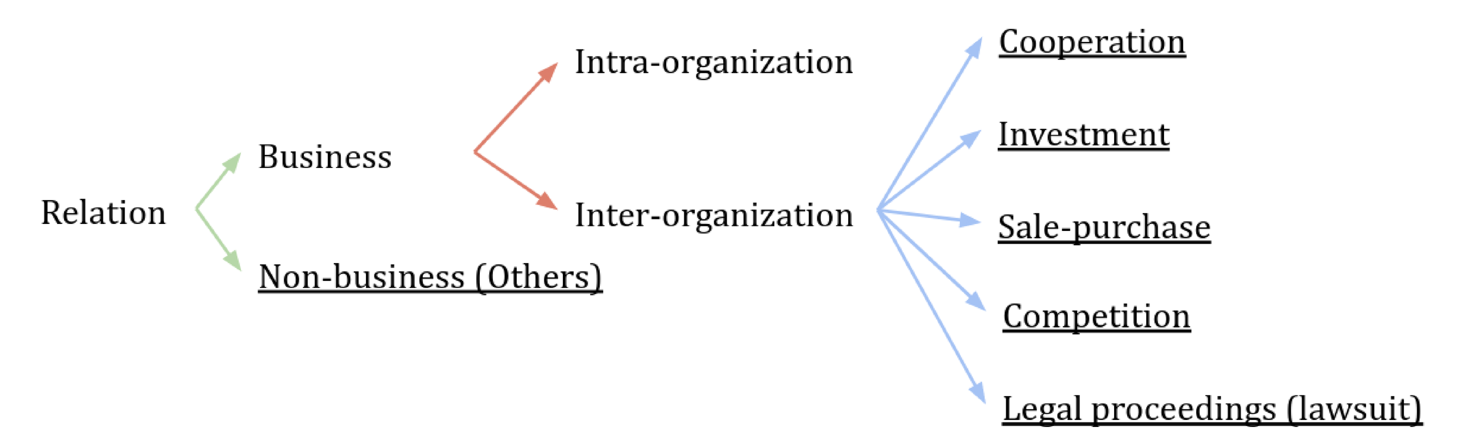

Figure 3. Characterizing Business Relations. (Relations we consider are underlined.)

Annotation Guideline

- Given a sentence S, and a set of entity pairs composed of non overlapping entities $\{(EO_1,\ EO_2), EO_i \in S\}$, annotation consists in assigning one relation R per entity pair among five business relations and one negative relation (cited earlier).
- It is important to note that many relation types can hold between a given entity pair in real world.
- Annotators are asked to only consider explicit mentions of relations in the current sentence *without any additional external knowledge*.
- $1k$ of data doubly annotated –> Cohen'K = 0.766 for English and Cohen'K = 0.685 for French

Annotation Results

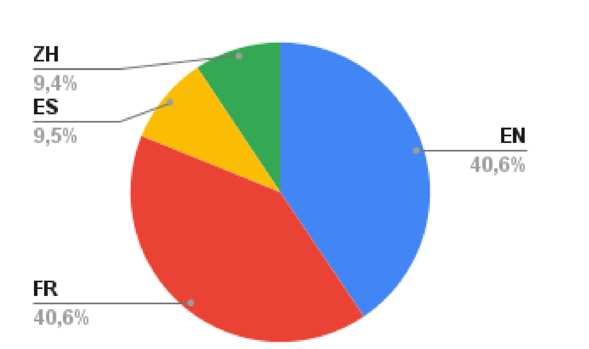

Figure 4. Data distribution per language.

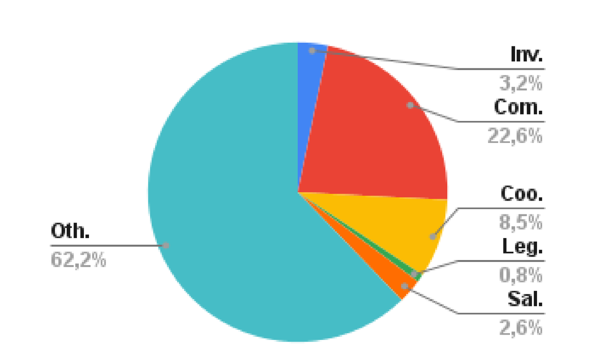

Figure 5. Data distribution per relation type.

Experimental Results

	Lang.	EN			FR			ES			ZH		
Settings	Models	P	R	F	P	R	F	P	R	F	P	R	F
S_0‡	Monolg.	67.7	71.9	69.5	72.2	66.8	69.0	74.4	72.5	73.1	**75.8**	**73.2**	**74.3**
S_1	EN	66.8	**72.4**	69.1	67.8	51.9	57.3	72.2	57.3	62.3	41.6	32.4	34.8
	FR	62.6	57.5	59.6	69.0	63.4	65.8	**78.3**	67.0	70.3	39.3	30.9	31.4
	ES	54.1	57.5	54.2	58.8	51.1	53.6	**77.1**	**76.8**	**76.8**	39.0	43.3	38.4
	ZH	49.6	32.3	35.5	50.7	29.4	32.7	54.5	36.4	40.7	62.9	72.2	66.0
S_2	$zero\_EN$	60.9	63.6	61.3	72.1	65.6	68.0	**83.3**	**<u>86.3</u>**	**<u>84.6</u>**	72.7	59.2	62.2
	$zero\_FR$	66.3	70.2	67.8	68.3	55.6	60.1	**<u>83.6</u>**	**79.1**	**81.0**	60.0	60.6	60.3
	$zero\_ES$	65.1	69.7	67.1	**73.1**	65.2	68.3	**79.6**	71.0	**73.9**	60.4	60.2	60.3
	$zero\_ZH$	66.9	70.8	68.3	**<u>74.4</u>**	**67.0**	**69.8**	**80.5**	**77.7**	**78.7**	60.3	52.0	54.2
S_3	$richL$	66.5	**<u>75.0</u>**	**<u>70.2</u>**	71.9	**67.5**	**69.4**	**77.8**	66.5	71.1	42.5	36.9	38.4
	$poorL$	58.6	56.7	56.9	61.5	50.7	54.8	**75.5**	**73.8**	**73.2**	53.7	54.3	54.0
S_4	all	**<u>67.8</u>**	**72.9**	**69.9**	**<u>74.4</u>**	**<u>68.8</u>**	**<u>70.8</u>**	**79.3**	**80.4**	**79.7**	73.8	64.1	65.1

Table 1. Monolingual and cross-lingual models results per language. Best performing models in each (S_i) setting are in bold while the best model for each language is underlined. ‡: Baselines models.

Findings

- (S_1) *Training on one language* does not necessarily transfer well to all other languages.
- (S_2) In *zero-shot* transfer, excluding ZH improves EN results, excluding EN or FR yield better results on ES.
- (S_3) Training on *poorly-labeled* data has weak transfer power compared to richly labeled data.
- (S_4) *All-joint transfer* that combines all languages during training was the best, beating all monolingual baselines.
- The relation types with the best F-score are the ones with more training data.
- Under-represented relation types gained an improvement over baseline models for many languages when training on more than one language.

References

[1] D. Braun, A. Faber, A. Hernandez-Mendez, and F. Matthes. Automatic relation extraction for building smart city ecosystems using dependency parsing. In P. Basile, V. Basile, D. Croce, F. Dell'Orletta, and M. Guerini, editors, *Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence (NL4AI 2018) co-located with 17th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2018), Trento, Italy, November 22nd to 23rd, 2018*, volume 2244 of *CEUR Workshop Proceedings*, pages 29–39. CEUR-WS.org, 2018.

[2] D. Burdick, M. Hernández, H. Ho, G. Koutrika, R. Krishnamurthy, L. C. Popa, I. Stanoi, S. Vaithyanathan, and S. R. Das. Extracting, linking and integrating data from public sources: A financial case study. *Available at SSRN 2666384*, 2015.

[3] S. Collovini, P. N. Gonçalves, G. Cavalheiro, J. Santos, and R. Vieira. Relation extraction for competitive intelligence. In *International Conference on Computational Processing of the Portuguese Language*, pages 249–258. Springer, 2020.

[4] D. De Los Reyes, A. Barcelos, R. Vieira, and I. Manssour. Related named entities classification in the economic-financial context. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 8–15, 2021.

[5] H. Khaldi, F. Benamara, A. Abdaoui, N. Aussenac-Gilles, and E. Kang. Multilevel entity-informed business relation extraction. In *International Conference on Applications of Natural Language to Information Systems*, pages 105–118. Springer, 2021.

[6] C. Yan, X. Fu, W. Wu, S. Lu, and J. Wu. Neural network based relation extraction of enterprises in credit risk management. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–6. IEEE, 2019.

[7] J. Zhao, P. Jin, and Y. Liu. Business relations in the web: Semantics and a case study. *Journal of Software*, 5(8):826–833, 2010.

[8] Z. Zuo, M. Loster, R. Krestel, and F. Naumann. Uncovering business relationships: Context-sensitive relationship extraction for difficult relationship types. In M. Leyer, editor, *Lernen, Wissen, Daten, Analysen (LWDA) Conference Proceedings, Rostock, Germany, September 11-13, 2017*, volume 1917 of *CEUR Workshop Proceedings*, page 271. CEUR-WS.org, 2017.

hadjer@geotrend.fr