How's Business Going Worldwide? A Multilingual Corpus for Business Relation Extraction
Khalid Hadjer 1,2, Benamara Farah 2, Pradel Camille 1, Sigel Grégoire 1, Aussenc-Gilhes 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 relying on Information Extraction techniques, such as business relation extraction (BRE) from web content. BRE aims at discovering either Inter-Organizational (Inter-ORG) relations linking a company and its components (e.g., company-employers, company-CEO) or Inter-Organizational (Inter-ORG) relations involving different companies (e.g., company-customer, company-partner).

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:

1. Collecting documents using a list of keywords related to various business activities.
2. Extracting named entities of type ORG.
3. Collecting sentences candidates for annotation.

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

Characterizing Business Relations

We study Inter-ORG relations in the following five categories:

- Business: e.g., "Acquire", "Merge", "Acquisition".
- Competition: e.g., "Competition", "Competition", "Joint Venture".
- Non-Business (Others): e.g., "Buy", "Sell".
- Cooperation: e.g., "Joint Venture", "Partnership".
- Legal proceedings: e.g., "Litigation".

Experimental Results

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

Findings

- [5] Training on one language does not necessarily transfer well to all other languages.
- [5] In zero-shot transfer, excluding ZH improves EN results, excluding EN or FR yield better results on ES.
- [5] Training on poorly labeled data has weak transfer power compared to richly labeled data.
- [5] All-lingual 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


