

A method for the extraction of relation patterns from document clusters

Description

The present invention relates to the technical field of natural language processing and in particular to the extraction of linguistic patterns expressing different types of relationship from natural language documents including news articles. Methods disclosed in the past (e.g. kernel or linear-pattern methods) all have important drawbacks such as laborious engineering tasks, need for human processing, slow relation pattern detection, very difficult output that is impossible to be manipulated by human experts, inherent limitations, and limited speed related to the use of public search engines.

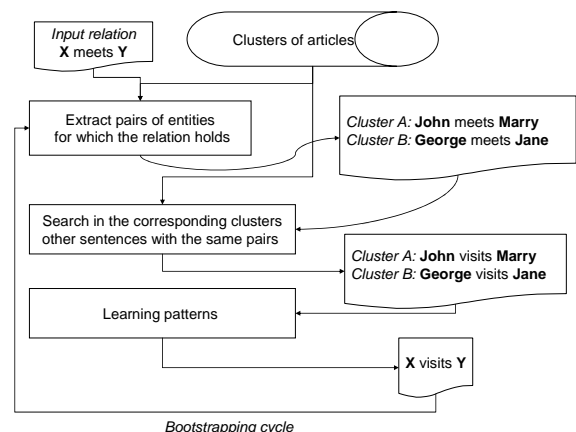
The proposed innovative method makes use of an automatic or semi-automatic procedure, to learn iteratively patterns for specific relations using clusters of similar articles. Pattern-based relation extraction is potentially faster than kernel methods which consider each pair of entities in the text. Moreover, pattern learning from article clusters is faster than similar pattern learning methods using the Web and allows the performance of numerous learning iterations which improves coverage. The use of clusters of articles also results in higher precision and more accurate classification into relationship types. Finally, the user can define the relations of interest, allowing the customisation of the method to meet specific user needs.

Innovative aspects and main advantages

- Relation extraction with learned patterns is faster than extraction based on Kernel methods
- Pattern learning from article clusters is faster than similar pattern learning processes using the Web
- Possibility for numerous learning iterations improving the coverage
- Higher precision and more accurate classification
- Human intervention possible in the learning process
- Algorithm can work in supervised or unsupervised modes
- Customisation of relations to meet user needs

Areas of application

- Analysis of news articles
- Learning of social networks
- Finding groups of related people/organizations and relations amongst named entities
- Helps find answers regarding specific relationships between people
- Automatic ontology building and database generation,
- Intelligent document searching and indexing



Stages of development

Patent priority date 16/04/07 (EP07106252.5)

Technology is mature and in continuous development. The European Media Monitor uses the technology. Non-exclusive licences are available.

Scientific contact

Hristo Tanev
IPSC – JRC - European Commission
I-21020 Ispra - Italy
Tel: +39 033278 6792
Email address: Hristo.tanev@jrc.it

Licensing contact

Intellectual Property and Scientific Collaboration Unit
JRC - European Commission
B-1049 Brussels, Belgium
Email: JRC-TechTransfer@ec.europa.eu
Reference: file n°2788