23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI'11), November 7-9 2011, Boca Raton, Florida (USA)

NASS: News Annotation Semantic System

Ángel Luis Garrido, Oscar Gómez (Ibercentro Media C & S, Spain)
Sergio Ilarri, Eduardo Mena (University of Zaragoza, Spain)

Problem

- News categorization in the Media using a **high number of thesaurus tags**
- Large number of items to categorize
- Few people in Documentation Departments

Newspaper Pages

Goal

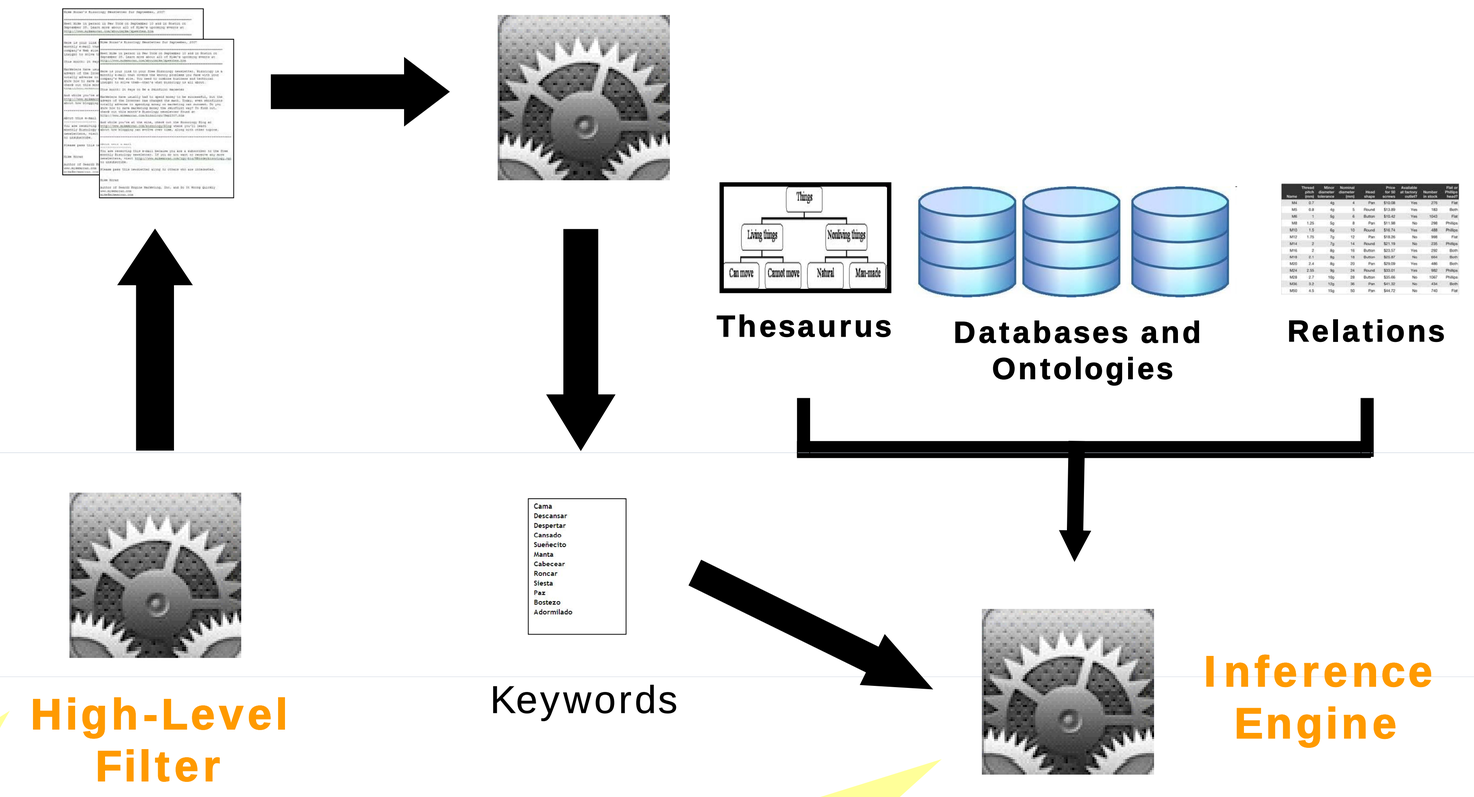
➔ Assisting in the daily work of categorization of news in a Documentation Department.

Advantages

- Combination of different techniques on different steps in order to optimize results
- Easy to manage and update
- High performance (99% accuracy)
- Money savings

Lemmatization, Named Entities and Obtention of Keywords

- Lemma: Canonical form of a word
- Standardization of terms
- Detection of proper names consisting of multiple words like *"New York"*
- Noise deletion
- Meaningless words discarding



SVM Filtering

- Advantages:
  - High speed
  - Good accuracy over general themes
- Drawbacks:
  - It needs training on every tag
  - Dichotomous nature
  - Low hit rates with concrete and changing themes (like is usual in the Media)

Ontologies and Heuristics

- Advantages:
  - Good accuracy on concrete topics
  - It handles a high number of tags
  - Easy to manage and update. Reusability
  - Semantic information and inference ability
- Drawbacks:
  - Low hit rates over general themes
  - It could be slow and hard to design when the number of entities is high

References

[1] D. C. Wimalasuriya and D. D., "Ontology-based information extraction: An introduction and a survey of current approaches," Journal of Information Science, vol. 36, no. 3, pp. 306–323, 2010.

[2] A. F. Smeaton, Using NLP or NLP Resources for Information Retrieval Tasks. Natural Language Information Retrieval. Kluwer Academic Publishers, 1999.

[3] S. Sekine and E. Ranchhod, Named Entities: Recognition, Classification and Use. John Benjamins, 2009.

[4] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in Proceedings of 10th European Conference on Machine Learning (ECML 98). Springer, pp. 137–142, 1998.

Acknowledgements

This work was supported by the CICYT project TIN2010-21387-C02-02 and IBERCENTRO MEDIA C.& S.

Contact

General information: **Angel Luis Garrido** (**algarrido@heraldo.es**)

Research group SID (*http://sid.cps.unizar.es*)