TALAD Project : a Coreference Resolution Chain for Nomination Detection

Mehdi Mirzapour, Michele Grimaldi *and* ***Jean-Yves Antoine, LIFAT, Tours University***

Motivation

To detect nominations in French social and political texts by resolving indirect coreference chains using Support Vector Machine trained by available French oral/written corpora.

General Strategy

•To use mention-paired classifier strategy among other possibilities such as clustering.

•Mentions referring to the same object would be included in a unique chain.

•For each anaphoric mention, a list of previous mentions occurring in a search window is created.

•Each pair receive a score and the candidate of the best pair is taken as antecedent.

•In this practice, we have only tested transitive closure of the coreferents pairs. No semantic constraint for the time being.

Features

Lexical: 14 features such as Number, Gender, Inclusion.

Syntactical: 7 features such as POS.

Semantical: 3 features such as Entities.

Distance Discourse: 3 features such as Mentions.

ANCOR Corpus

Corpus	Speech type	Interactivity	Size	Duration
ESLO	interview	low	452,000 words	27,5 hours
ESLO_ANCOR			*417,000 words*	*25 hours*
ESLO_CO2			*35,000 words*	*2.5 hours*
OTG	task oriented conversational speech	high	26,000 words	2 hours
Accueil_UBS	phone conversational speech	high	10,000 words	1 hour

Possible Feature Works

- **We have made a F1 recall of 0.92 for OTG sub corpus and we are going to extend it very soon to all of the other sub corpora.**
- **We will have hybrid/single training with written French DEMOCRAT corpus as well and compare the result together.**
- **We want to have more semantical relations than what we currently have. We may use embeddings coming from RezoJDM graph or may be from other language models like BERT families or fasttext.**

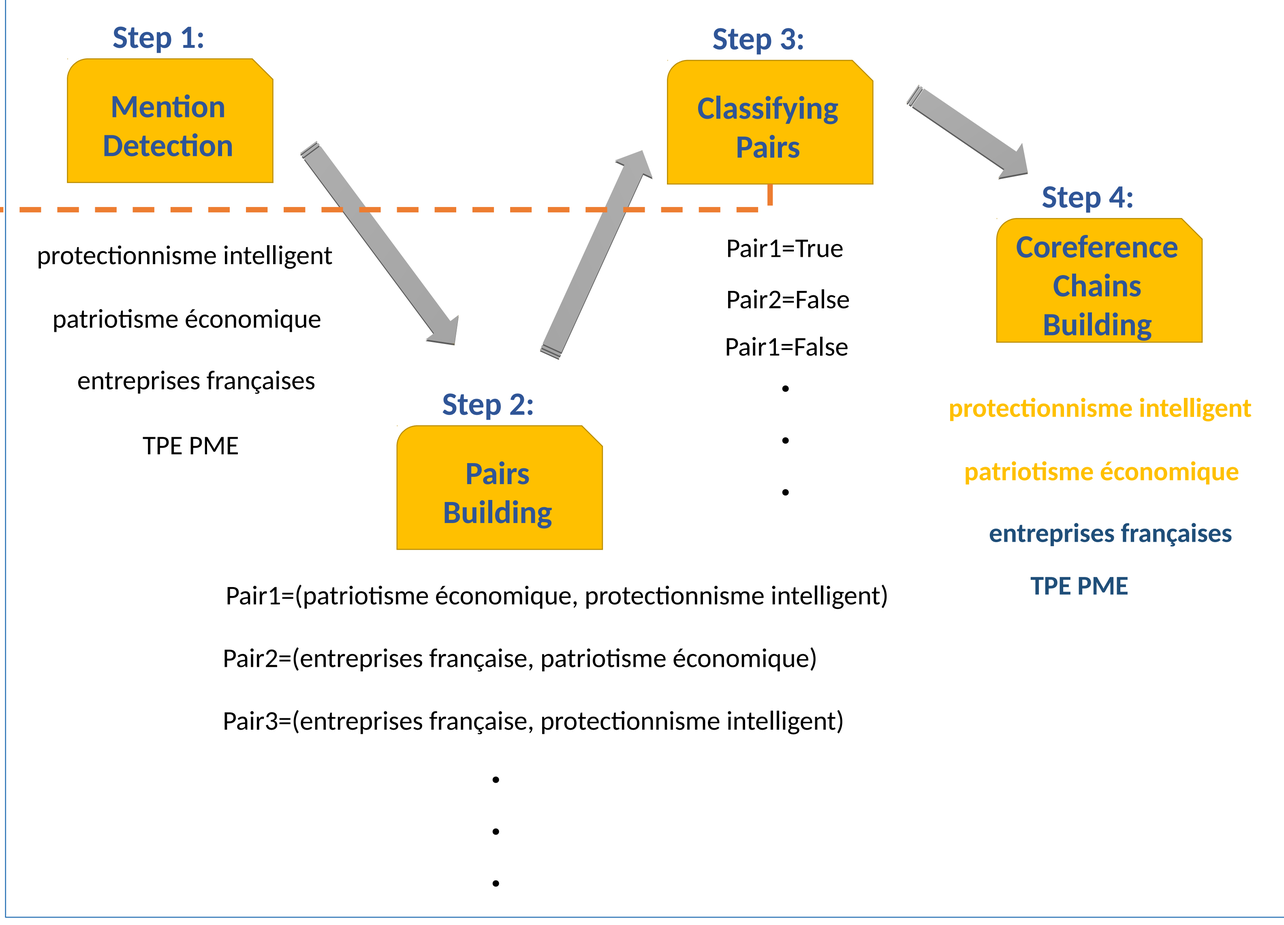
Input

•The input to the system is a text in multiple lines.

Example:

… à mettre en œuvre du protectionnisme intelligent à mettre en avant du patriotisme économique pour donner un avantage aux entreprises françaises dans la commande publique voilà tout cela. Le patriotisme économique qui n'a jamais été mis en œuvre le protectionnisme intelligent la défiscalisation des heures supplémentaires la suppression du travail détaché la baisse des charges mais exclusivement pour les TPE PME.

Pipe Lines



Output

•The output is the resolved coreference chains.

Example:

… à mettre en œuvre du **protectionnisme intelligent** à mettre en avant du **patriotisme économique** pour donner un avantage aux **entreprises françaises** dans la commande publique voilà tout cela. Le **patriotisme économique qui** n'a jamais été mis en œuvre le **protectionnisme intelligent** la défiscalisation des heures supplémentaires la suppression du travail détaché la baisse des charges mais exclusivement pour les **TPE PME**.

References

1. Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., Antoine, J. Y., & Dinarelli, M. (2016, April). Coreference Resolution for French Oral Data: Machine Learning Experiments with ANCOR. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 507-519). Springer, Cham.
2. Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.Y., Pelletier, A., Maurel, D., Eshkol, I., Villaneau, J.: Ancor centre, a large free spoken french coreference corpus: description of the resource and reliability measures. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. Reykjavik, Iceland (2014)