Combination of Contextualized and Non-Contextualized Layers for Lexical Substitution in French

Kévin Espasa[1], Emmanuel Morin[1], Olivier Hamon[2]
[1] *LS2N, 2 Chemin de la Houssinière, 44322 Nantes, France*
[2] *Syllabs, 35-37 rue Chanzy, 75011 Paris, France*

Introduction

— Lexical substitution (LS) task requires to substitute a target word by candidates in a given context. Candidates must keep meaning and grammatically of the sentence.

— LS have two steps: find a list of substitutions for the target word and then rank them in order to find the best substitute in the sentence context.

— We propose an application of the state-of-the-art method based on BERT in French and a novel method using contextualized and non-contextualized layers

Example

Sentence: *Benzema a **marqué** un but.* (Benzema scored a goal.)

Target word: *marqué* (scored)

Substitute candidates from WOLF: *inscrit* (scored), *coché* (mark)

✓ *Benzema a **inscrit** un but.* (Benzema scored a goal.)

✗ *Benzema a **coché** un but.* (Benzema marked a goal.)

State-of-the-art

Propositions of Substitute Candidates:

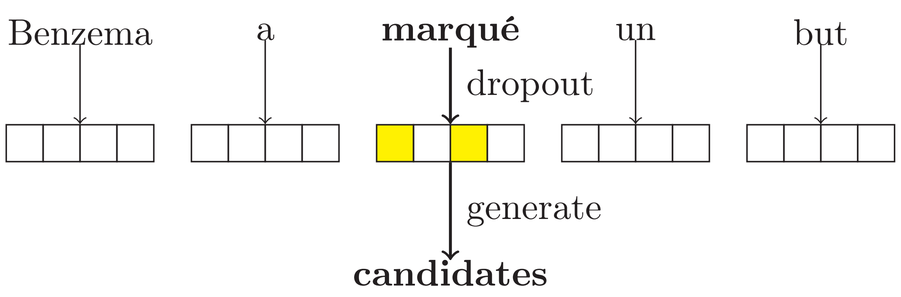
- Lexical resources (WordNet, dictionnaries or both)
- Vector space models
- Pretrained language models

Ranking substitute candidates:

- Classifiers
- Vector space models
- Pretrained language models

Our method

1. Propositions of Substitute Candidates:



2. Ranking substitute candidates:

- Similarity in the first and last layer which include non-contextualized and contextualized information for the target word and a candidates
- Similarity foreach other word in sentence in order to compare the impact of the substitution on the meaning and choose the best candidate.

3. Post process:
Substitution words are lemmatized in the gold standard. We make two steps of post process to match with them:

- Lemmatization (inscrit→ inscrire)
- Remove duplicated candidates

Evaluation Tasks

Evaluation tasks	SemDis 2014 V1	SemDis 2014 V2
#Sentences	300	285
#Unique target words	10	10
#Candidates	1,771	6,034

Example of sentence in second gold standard:
*90 % de ces hommes ont été **arrêtés** pour des délits liés à la drogue*
Substitution words from gold standard for this sentence:
coffrer ; appréhender ; inculper ; écrouer ; emprisonner

Evaluation measures

$$best_{norm}(i) = \frac{score_i(best_i)}{score_i(max_i)} \quad (1)$$

$$oot_{norm}(i) = \frac{\sum_{b \epsilon P_i} score_i(b)}{\sum_{a \epsilon M_i \subset G_i} score_i(a)} \quad (2)$$

$score_i(candidate)$: The score of candidate in the gold standard
$best_i$: Best candidate proposed by the system
P_i: List of candidates proposed by the system
max_i: Best candidate from the gold standard
M_i: the subset of G_i which contains the 10 best scores

Evaluations

Method	BEST	OOT
Desalle et al. (2014)	**0.29**	**0.41**
Zhou et al. (2019)	**0.29**	0.31
Our method	0.25	0.32
Ferret (2014)	0.23	0.29
Gábor (2014)	0.17	0.22
Fabre et al. (2014)	0.13	0.33

Evaluation on first gold standard

Method	BEST	OOT
Desalle et al. (2014)	**0.48**	**0.38**
Ferret (2014)	0.33	0.33
Our method	0.30	0.24
Zhou et al. (2019)	0.30	0.23
Gábor (2014)	0.29	0.19
Fabre et al. (2014)	0.17	0.28

Evaluation on second gold standard

Conclusions

— Our method increases on the OOT but decreases on the BEST measure in the SemDis 2014 benchmark in comparison with Zhou et al. (2019) but do not have a better score than (Desalle et al., 2014)

— We want to applied our method in English to have better comparison with the BERT-based method (Zhou et al., 2019)

— A method that can consider a multi-word like a target and replace this with a single word substitute or a multi word substitute.

Bibliographical References

Desalle, Y., Navarro, E., Chudy, Y., Magistry, P., and Gaume, B. (2014). BACANAL: Short length random walks for lexical analysis, application to lexical substitution (BACANAL : Balades aléatoires courtes pour ANAlyses lexicales application à la substitution lexicale) [in French]. In *TALN-RECITAL 2014 Workshop SemDis 2014 : Enjeux actuels de la sémantique distributionnelle (SemDis 2014: Current Challenges in Distributional Semantics)*, pages 206–217, Marseille, France. Association pour le Traitement Automatique des Langues.

Fabre, C., Hathout, N., Ho-Dac, L.-M., Morlane-Hondère, F., Muller, P., Sajous, F., Tanguy, L., and Van de Cruys, T. (2014). TALN-RECITAL 2014 workshop SemDis 2014 : Enjeux actuels de la sémantique distributionnelle (SemDis 2014: Current challenges in distributional semantics). Marseille, France. Association pour le Traitement Automatique des Langues.

Ferret, O. (2014). Using a generic neural model for lexical substitution (utiliser un modèle neuronal générique pour la substitution lexicale) [in French]. In *TALN-RECITAL 2014 Workshop SemDis 2014 : Enjeux actuels de la sémantique distributionnelle (SemDis 2014: Current Challenges in Distributional Semantics)*, pages 218–227, Marseille, France. Association pour le Traitement Automatique des Langues.

Gábor, K. (2014). The WoDiS system - WOlf and DIStributions for lexical substitution (le système WoDiS - WOLF et DIStributions pour la substitution lexicale) [in French]. In *TALN-RECITAL 2014 Workshop SemDis 2014 : Enjeux actuels de la sémantique distributionnelle (SemDis 2014: Current Challenges in Distributional Semantics)*, pages 228–237, Marseille, France. Association pour le Traitement Automatique des Langues.

Zhou, W., Ge, T., Xu, K., Wei, F., and Zhou, M. (2019). BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.