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

Kévin Espasa¹, Emmanuel Morin¹, Olivier Hamon²
(1) LS2N Nantes University, France
(2) Syllabs, Paris, France
{firstname.name}@univ-nantes.fr
{name}@syllabs.com

Abstract

Lexical substitution task requires to substitute a target word by candidates in a given context. Candidates must keep meaning and grammatically of the sentence. The task, introduced in the SemEval 2007, has two objectives. The first objective is to find a list of substitutes for a target word. This list of substitutes can be obtained with lexical resources like WordNet or generated with a pre-trained language model. The second objective is to rank these substitutes using the context of the sentence. Most of the methods use vector space models or more recently embeddings to rank substitutes. Embedding methods use high contextualized representation. This representation can be over contextualized and in this way overlook good substitute candidates which are more similar on non-contextualized layers. SemDis 2014 introduced the lexical substitution task in French. 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 to increase the suggestion of words having a lower probability in a given context but that are more semantically similar. Experiments show our method increases the BERT based system on the OOT measure but decreases on the BEST measure in the SemDis 2014 benchmark.

Keywords: Lexical substitution, French language, Combination contextualized layers

1. Introduction

Lexical substitution (McCarthy and Navigli, 2007) is the task of proposing substitution words to replace a target word in a given context. The task can be split into two issues. Firstly, how to find a list of candidates (or synonyms) for the target word and secondly how to validate or invalidate them in a given textual context. For instance, “She paints the bank of the river”, edge can substitute bank while keeping the meaning and the grammaticality of the sentence. Whereas the substitution by the synonym financial institution modifies the meaning of bank. The ranking methods use the context of the sentence to understand the sense of bank to accept edge and reject financial institution.

Historically, the list of synonyms for a target word is obtained in lexical resources which contain synonyms like WordNet (Miller, 1992). After picking, candidates are ranked by their quality to replace the target word in its context. With the language models rise using contextual embeddings like BERT, ELMO, candidates are generated by the model with the contextual information. Then, semantic similarity, between target and candidate of the original sentence and the sentence containing the substitution, is computed to compare the impact of the substitution on the meaning and choose the best candidate.

Lexical substitution is used in many applications like semantic expansion (Han et al., 2020), paraphrasing (Thater et al., 2009), word sense induction (Alagić et al., 2018) or data augmentation (Xiang et al., 2021).

The main contributions of this paper are:

- Application of state-of-the-art method on French and comparison with previous methods applied on the French evaluation corpus SemDis 2014.
- A novel method using a combination of low contextualized layer and high contextualized layer in order to increase the score of words which are not present repeatedly in a given context but highly similar to the target word.

2. Related Work

In the original lexical substitution task defined by McCarthy and Navigli (2007), systems have to pick or generate candidates from lexical resources and then validate them with ranking methods using a context disambiguation. In this section, we present related work for the two components: substitute candidates propositions and ranking methods. We also present the evaluation campaign in English and French for this task.

2.1. Propositions of Substitute Candidates

The objective of first component of lexical substitution systems is to find for a target word in a given context a list of new words who can substitute it.

Three strategies are used for this. Lexical resources strategy use WordNet (Gábor, 2014), dictionaries (Ferret, 2014) Desalle et al., 2014 or combination of both (Hassan et al., 2007) to find a list of synonyms for the target word. This strategy is highly impacted by the quality of lexical resources. A good substitute in a given a context can be missing within the resources.
For example, in the sentence “Benzema is the forward of Real Madrid” a good candidate for the replacement can be player but WordNet or its French equivalent WOLF (Sagot and Fiser, 2008) do not contain this word in the list of synonyms.

The strategy using vector space models generates candidates using word embedding. Instead of using lexical resources, this methods generate candidates with a Word2Vec model (Mikolov et al., 2013) trained on a huge corpora. Models propose candidates using similarity between them and the target word in vector space (Melamud et al., 2015; Roller and Erk, 2016). Melamud et al. (2016) extend this idea with Context2Vec. Focus of word representation is not only on the target word but on the entire context representation.

The last strategy uses large pretrained language models like BERT (Devlin et al., 2019) to predict a word in a given context. Zhou et al. (2019) propose to partially mask the target word in BERT in order to give to the model a few pieces of information to guide the propositions. Arefyev et al. (2020) uses different language models like XLNET, BERT and ELMO to generate a list of candidates.

2.2. Ranking Candidate

The second component of lexical substitution systems aims to rank the candidates in relation to their similarity with the target word in a given context. Different approaches are used to classify candidates based on substitution quality in a given context. All of the methods described bellow compute the similarity between target and candidate with contextualization of the word in the sentence.

Hassan et al. (2007) propose a combination of lexical, semantic and probabilistic features to compute the similarity. Szarvas et al. (2013) train a supervised Max-Entropy classifier on a delexicalisation features like local n-grams frequencies, number of synsets in WordNet. Graph based methods using short random walks or directional similarity are also used to rank candidates (Desalle et al., 2014). Vector space modeling using contextual representations (Thater et al., 2015; Dinu and Lapata, 2010; Gábor, 2014), word and context embedding are used to compute the similarity between the target and the candidate (Ferret, 2014; Melamud et al., 2015; Roller and Erk, 2016).

Recently, methods using pre-trained language models compute similarity using representation in contextual embedding of word and target. Zhou et al. (2019) compute similarity for each word embedding between two sentences: the original sentence and the candidate sentence using BERT. Arefyev et al. (2020) uses different models: ELMO (Peters et al., 2018), BERT and XLNET (Yang et al., 2020) with different methods to compute the similarity using embedding or dynamic patterns.

There are two possibilities to retrieve the candidates: picking from lexical resources or generating with language models. Lexical resources depend on the quality of content, some words with specific relation like hypernym could be overlooked. Language models’ key limitation is the probability to generate a word not linked to the target word. Ranking methods use contextual information to highlight certain candidates over others. Pre-trained language models could be too contextualized in high layers. For example: “Des olives et des avocats y poussent” (Olives and avocados grow there.), a good synonym for avocats (avocados) is avocatiers (avocado trees) but in the last and most contextualized layer of BERT. amandes (almonds) is more similar to avocats than avocatier in the given context.

3. Experiments

We are not aware of any work applying pre-trained language models on French for lexical substitution task. We focus our experiments on pretrained language models, in particular on state-of-the-art method. We also propose a novel method using the difference in the level of contextualization between first and last layers. We use CamemBERT (Martin et al., 2020), the French state-of-the-art language model based on RoBERTa architecture (Liu et al., 2019). In this section, we present our hypothesis, then the BERT based model system and finally the application of our hypothesis. We also describe the French datasets that we use to evaluate our methods.

3.1. Hypothesis

Camembert is a bidirectional Transformer encoder trained on Oscar corpus (Ortiz Suárez et al., 2019) with a masked language modeling objective. Contextual language models like CamemBERT have two advantages for the lexical substitution task: (i) to generate a candidate with information from left and right context, (ii) to compute semantic similarity between the original sentence (with target word) and the sentence containing the candidate. Only one model, without other resources, can respond to the lexical substitution components which are generating candidates and ranking them.

Zhou et al. (2019) use the last 4 layers in BERT to compute the impact of candidates in the sentence. These layers are the most contextualized and validate if the injection of a candidate in a sentence does not change the overall meaning of the sentence. The main limitation is the length of a sentence could be negatively impacted by the similarity between two words in these layers.

To illustrate our limitation, we present 4 sentences with the same target word: avocats (avocados). For each sentence, the meaning of the target does not change. We only add more context which does not impact the representation of avocats.

Sent 1: Des avocats y poussent.

(Avocados grow there.)
Sent 2: Des olives et des avocats y poussent. (Olives and avocados grow there.)

Sent 3: Des oranges, des olives et des avocats y poussent. (Oranges, olives and avocados grow here.)

Sent 4: Cette région bénéficie d’un microclimat, ce qui fait que des oranges, des olives et des avocats y poussent. (This region benefits from a microclimate, which causes oranges, olives and avocados to grow here.)

Table 1: Evolution of cosine similarity $[-1, 1]$ between the French word avocats and candidates

| Candidate  | Layer | Sent 1 | Sent 2 | Sent 3 | Sent 4 |
|------------|-------|--------|--------|--------|--------|
| avocatiers (avocado trees) | 1     | 0.50   | 0.50   | 0.50   | 0.50   |
|           | 12    | 0.70   | 0.64   | 0.64   | 0.58   |
| amandes (almonds)             | 1     | 0.38   | 0.38   | 0.38   | 0.38   |
|           | 12    | 0.67   | 0.70   | 0.75   | 0.82   |
| fromages (cheeses)            | 1     | 0.27   | 0.26   | 0.26   | 0.26   |
|           | 12    | 0.39   | 0.47   | 0.65   | 0.64   |

After the candidate generations, a method based on influence of substitution on a given context, ranks them. Model generated words are not always good substitutes. The goal of the ranking method is to validate or invalidate a candidate using influence on the other word embedding in the sentence. For each token presents in two sentences, sentence with target word and sentence with replacement of target word in position $k$ by a candidate, influences are calculated with cosine similarity between two embeddings.

$$s_v(x_k'|x, k) = \sum_i w_{i,k} \times \Lambda(h(x_i'|x), h(x_i'|x'))$$ (1)

Where $x$ is the sentence with target word and $x'$ the sentence with replacement of the target by the candidate at position $k$. $\Lambda(h(x_i|x), h(x_i'|x'))$ is the cosine similarity between token representation on the last four layers at position $i$ in the sentence $x$ and in the sentence $x'$. $W_{i,k}$ is used to weight each token with their semantic dependencies. Weights are calculated using the average of self-attention from $i^{th}$ token to $k^{th}$ position in $x$. A proposition score is also calculated:

$$s_p(x_k'|x, k) = \log\frac{P(x_k'|\tilde{x}, k)}{P(x_k|\tilde{x}, k)}$$ (2)

Where $x_k'$ is the candidate, $x$ the sentence and $k$ the position in the sentence of target. $\tilde{x}$ is the sentence with the dropout applied to the target.

These two equations are used to assign a score to a candidate in the given context.

$$s(x_k'|x, k) = s_v(x_k'|x, k) + \alpha \times s_p(x_k'|x, k)$$ (3)

Where $\alpha$ is a weight. This method uses only highly contextualized layers to compute the similarity between target word and candidate. This contextualization could have a negative impact on the similarity between two words. The target word could become overcontextualized and loss in similarity with words yet closer semantically. We propose a method that keeps the influence score ($s_v$) but replaces the propositional score ($s_p$) by a similarity in the first and last layer which include non-contextualized and contextualized information.

3.2. Bert Based Lexical Substitution

Zhou et al. (2019) propose a method to generate and rank candidates using BERT. Rather than masking the target word in order to generate candidates which are semantically different, a dropout is applied on the embedding of the target word. The idea behind this is to give partial information to the language model. As a part of embedding is randomly masked, the model suggests the closest candidates to the target word. If the value of the dropout is too high, then the model proposes the target word, but if the value of the dropout is too low, then the model proposes candidates which are too semantically different.

3.3. Ranking with First and Last Embedding

As described previously, last layers are more impacted by context. Some random words could have a greater cosine similarity with the target word than a good candidate. As shown in Table[1] the first layer is less impacted by the context. We propose to combine the similarity between the first and the last layer in order to rank the list of candidates. The objective of the first layer is to improve the score when the candidate is close to the target without context. The last layer aims to validate if this candidate is suitable for the global context.

First, we generate our list of candidates with CamemBERT. The sentence with the target at position $k$ is encoded. We apply the dropout method on the embedding
Figure 1: Example of workflow for the sentence *Benzema a gagné le championnat* (*Benzema won the championship*) with the target word *gagné* (won). Generated method proposes a list of candidates, then we use the ranking method. Finally, he last step aims to filter candidates in order to remove duplicate candidates like gagné (won) and gagner (win) and to have lemmas to match with gold standard.
In order to respect the expected format for evaluation, we lemmatize candidates generated by CamemBERT with Spacy\footnote{https://spacy.io/}. We also remove duplicated candidates using their lemma, the language model can suggest the same word with different gender and number agreement or typology. Finally, we remove a candidate for which its lemma is the same as the target word.

Regarding the BERT-based method and our method, we experiment different parameters. We try with two values for the dropout: 0.1 and 0.3. The global score described by \cite{zhou2019bert} have a parameter named alpha in the equation\footnote{https://spacy.io/} authors tried different values and choose 0.1. We use the same value to reproduce this method. In our own equations, we also have two values. The first parameter is alpha in equation\footnote{https://spacy.io/} for which we propose two values 0.1 and 0.01. The first value gives a better importance to our layer score. The second parameter is beta in equation\footnote{https://spacy.io/} that it gives more or less importance to the similarity of the first layer or the last one. We tried a different beta between 0.1 and 0.9 with a step of 0.1.

The Table\footnote{https://spacy.io/} shows the result of our different parameters in comparison with the BERT-based method and the validation method ($s_v$) defined by \cite{zhou2019bert}. These results are evaluated with the first gold standard. The dropout value 0.1 gives better scores for BEST and OOT than the value to 0.3. The target word is a little masked, language model have less information about the target word, so CamemBERT suggests substitutes with greater semantic similarity. With the alpha to 0.01, the increase of beta seems to have a positive effect on both metrics. When beta is 0.8 or 0.9, the score is higher than the $s_v$ method but still lower than \cite{zhou2019bert} method including the propositional score. With the alpha to 0.1, the BEST metric for both dropout values decreases when the beta parameter increases. With the maximum value for beta, BEST score loses 0.6 compared to \cite{zhou2019bert}. As for the alpha at 0.01, the OOT metric increases with the augmentation of beta. From a beta to 0.5 on dropout to 0.1, the score is equal to \cite{zhou2019bert} and increases for each step. For both dropouts, OOT has a better score with the maximum value of beta.

The Table\footnote{https://spacy.io/} shows the evaluation of different methods and parameters on the second gold standard. Globally, between the two gold standards, the BEST score is higher with the second while the OOT decreases by around 0.07 point. A reason could be that the number of substitutes for each sentence has more than doubled (an average of 7 substitutes for the first gold standard and 16 for the second). The alpha value seems to have a weak impact on metrics mainly because the beta parameter does not only decrease the performance of the system when the first layer has more weight in the layer score. Contrary to the first gold standard, the beta value does not continually improve the score on OOT, from 0.6 the progression is less high.

In comparison to the \cite{zhou2019bert} method, the contribution of a non-contextualized layer has a nega-

| sentence | Benzema a gagné le championnat (Benzema won the championship) |
|----------|---------------------------------------------------------------|
| dropout 0.1 | gagné, remporté, joué (won, won, played) |
| dropout 0.5 | gagner, perdu, fait (win, lose, done) |
| dropout 0.9 | fait, commencé, rendu (done, started, rendered) |

Table 2: Influence of the dropout value on the suggestion made by the system for the target word gagné.
| Method                  | BEST  | OOT  | BEST  | OOT  |
|------------------------|-------|------|-------|------|
|                        | $d = 0.1$ | $d = 0.1$ | $d = 0.3$ | $d = 0.3$ |
| Zhou et al. (2019)     | 0.291 | 0.315 | 0.274 | 0.273 |
| $S_v$                  | 0.276 | 0.307 | 0.281 | 0.260 |
| $\alpha = 0.1; \beta = 0.1$ | 0.269 | 0.298 | 0.264 | 0.262 |
| $\alpha = 0.1; \beta = 0.2$ | 0.265 | 0.304 | 0.258 | 0.265 |
| $\alpha = 0.1; \beta = 0.3$ | 0.262 | 0.309 | 0.252 | 0.273 |
| $\alpha = 0.1; \beta = 0.4$ | 0.273 | 0.310 | 0.257 | 0.278 |
| $\alpha = 0.1; \beta = 0.5$ | 0.264 | 0.315 | 0.250 | 0.283 |
| $\alpha = 0.1; \beta = 0.6$ | 0.251 | 0.322 | 0.247 | 0.287 |
| $\alpha = 0.1; \beta = 0.7$ | 0.244 | 0.327 | 0.225 | 0.289 |
| $\alpha = 0.1; \beta = 0.8$ | 0.242 | 0.327 | 0.221 | 0.289 |
| $\alpha = 0.1; \beta = 0.9$ | 0.231 | 0.330 | 0.214 | 0.294 |
| $\alpha = 0.01; \beta = 0.1$ | 0.273 | 0.306 | 0.274 | 0.265 |
| $\alpha = 0.01; \beta = 0.2$ | 0.272 | 0.307 | 0.274 | 0.265 |
| $\alpha = 0.01; \beta = 0.3$ | 0.270 | 0.308 | 0.275 | 0.266 |
| $\alpha = 0.01; \beta = 0.4$ | 0.274 | 0.308 | 0.273 | 0.270 |
| $\alpha = 0.01; \beta = 0.5$ | 0.273 | 0.308 | 0.279 | 0.270 |
| $\alpha = 0.01; \beta = 0.6$ | 0.273 | 0.311 | 0.272 | 0.270 |
| $\alpha = 0.01; \beta = 0.7$ | 0.274 | 0.310 | 0.272 | 0.276 |
| $\alpha = 0.01; \beta = 0.8$ | 0.279 | 0.312 | 0.276 | 0.278 |
| $\alpha = 0.01; \beta = 0.9$ | 0.282 | 0.314 | 0.273 | 0.279 |

Table 3: BEST [0, 1] and OOT [0, 1] scores with different configuration for the $\alpha$ parameter in equation 5 and the $\beta$ parameter in the equation 6 evaluate with the first gold standard. $d$ is the dropout value on the target’s embedding.

Table 4: BEST [0, 1] and OOT [0, 1] scores with different configuration for the $\alpha$ parameter in equation 5 and the $\beta$ parameter in the equation 6 evaluate with the second gold standard.

Table 5: Normalized BEST [0, 1] and OOT [0, 1] scores for systems evaluated on first gold standard.

About the SemDis 2014 evaluation task, 3 participants submit between 1 and 5 methods. In order to clarify the similarity in a high level of contextualisation but semantic close without context increase their score. About the SemDis 2014 evaluation task, 3 participants submit between 1 and 5 methods. In order to clarify the table of results, we only keep the best submission of each team. A baseline system is also released by Fabre et al. (2014). We describe below the best proposal for each team and the baseline.

- Desalle et al. (2014) propose a method using random walks on a graph constructed from lexical resources JeudeMots\(^2\) and DicoSyn\(^3\).
- Ferret (2014) uses the cosine similarity between substitute picking in the dictionary Word XP and all words (except stopwords and the target word) in the sentence.
- Gábor (2014) uses WOLF and a vector representation to classify substitutes.
- Fabre et al. (2014) propose the baseline by picking in the dictionary DicoSyn a list of candidates for a target word and rank them using their frequency in the FRWAC (Baroni et al., 2009).

The Table 5 compares Zhou et al. (2019)’s method and our method with dropout to 0.1, alpha to 0.1 and beta to 0.6 and the other methods on the first gold standard. The Table 5 presents the same systems and parameters but evaluated on the second gold standard. On both gold standard, Desalle et al. (2014) obtains a better score, the gap is more important on the second gold standard. The baseline without using context has a better OOT score in both evaluations. This suggests that substitutes generated by CamemBERT were not performant. To support our words, the language model is unable to suggest a single real word in 39 sentences. Some words such as épécher (peel), essuyer (wipe), faucher (mow) and vaseux (muddy) could be problematic to the model, it generates candidates like ép, es, fu which do not exist in French.

\(^2\)www.jeudemots.org
\(^3\)www.cnrtl.fr/synonymie/

\[\begin{array}{|c|c|c|c|}
\hline
\text{Method} & \text{BEST} & \text{OOT} \\
\hline
\text{Desalle et al. (2014)} & 0.29 & 0.41 \\
\text{Zhou et al. (2019)} & 0.29 & 0.31 \\
\alpha = 0.1; \beta = 0.6 & 0.25 & 0.32 \\
\text{Ferret (2014)} & 0.23 & 0.29 \\
\text{Gábor (2014)} & 0.17 & 0.22 \\
\text{Fabre et al. (2014)} & 0.13 & 0.33 \\
\end{array}\]
Table 6: Normalized BEST [0, 1] and OOT [0.1] scores for systems evaluated on second gold standard

| Method                  | BEST | OOT |
|-------------------------|------|-----|
| (Desalle et al., 2014)  | 0.48 | 0.38|
| (Ferret, 2014)          | 0.33 | 0.33|
| $\alpha = 0.1; \beta = 0.6$ | 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|

5. Conclusions and Future Work

In this work, we propose an application of the state-of-the-art method in French. This method can suggest substitutes and rank them using the influence of the candidate on the sentence context. We propose a novel method, which combines the influence score from the Zhou et al. (2019) method and adds a layer score. The objective of this score is to increase the global score of a candidate which is less present in a given context but semantically closer to the target word on the first layer. With more non-contextualized information, the system outperforms the Zhou et al. (2019) method on the OOT, which evaluates the top 10 substitutes, at the expense of the BEST metric.

The performance of the system on the best candidate is the main limitation of our method. We have a better score on OOT, so we can conclude that our method is efficient on the selection of the 10 best candidates. Therefore, we can use this to select the top 10 and then use another method to rank this top 10.

Another point of improvement is the management of multi-words. We want to propose a method that can consider a multi-word like a target and replace this with a single word substitute or a multi word substitute. Masked language models could not suggest multi-words but it could be used in order to rank the list of candidates.

6. Bibliographical References

Alagić, D., Šnaider, J., and Padó, S. (2018). Leveraging lexical substitutes for unsupervised word sense induction. August.

Arefyev, N., Sheludko, B., Podolskiy, A., and Panchenko, A. (2020). A comparative study of lexical substitution approaches based on neural language models.

Baroni, M., Bernardini, S., Ferresi, A., and Zanchetta, E. (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation, 43:209–226, 09.

Desalle, Y., Navarro, E., Chudy, Y., Magistry, P., and Gaume, B. (2014). BACANAL: Short 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, July. Association pour le Traitement Automatique des Langues.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dinu, G. and Lapata, M. (2010). Measuring distributional similarity in context. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1162–1172, Cambridge, MA, USA, October.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 55–65, Hong Kong, China, November. Association for Computational Linguistics.

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, July. 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, jul. 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, jul. Association pour le Traitement Automatique des Langues.

Han, J., Sun, A., Zhang, H., Li, C., and Shi, S. (2020). Case: Context-aware semantic expansion. AAAI Conference on Artificial Intelligence, 34(05):7871–7878, apr.

Hassan, S., Csomai, A., Banea, C., Sinha, R., and Mihalcea, R. (2007). UNT: SubFinder: Combining knowledge sources for automatic lexical substitution. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 410–413, Prague, Czech Republic, June. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, D., and Sagot, B. (2020). Camembert: a tasty french language model. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.

McCarthy, D. andNavigli, R. (2007). SemEval-2007 task 10: English lexical substitution task. In Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval ‘07, page 48–53. Association for Computational Linguistics.

Melamud, O., Levy, O., and Dagan, I. (2015). A simple word embedding model for lexical substitution. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pages 1–7, Denver, CO, USA, June. Association for Computational Linguistics.

Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional LSTM. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 51–61, Berlin, Germany, August. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Miller, G. A. (1992). WordNet: A lexical database for English. In Speech and Natural Language: Proceedings of a Workshop Held, Harriman, NY, USA, oct.

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations.

Roller, S. andErik, K. (2016). PIC a different word: A simple model for lexical substitution in context. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1121–1126, San Diego, CA, USA, June. Association for Computational Linguistics.

Sagot, B. andFišer, D. (2008). Building a free French wordnet from multilingual resources. In OntoLex, Marrakech, Morocoo, may.

Szavas, G., Biemann, C., andGurevych, I. (2013). Supervised all-words lexical substitution using delexicalized features. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1131–1141, Atlanta, GA, USA, June. Association for Computational Linguistics.

Tanguy, L., Fabre, C., andRivière, L. (2018). Extending the Gold Standard for a Lexical Substitution Task: is it worth it? In LREC, Miyazaki, Japan, may.

Thater, S., Dinu, G., andPinkal, M. (2009). Ranking paraphrases in context. In Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer), pages 44–47, Suntec, Singapore, August. Association for Computational Linguistics.

Thater, S., Fürstenau, H., andPinkal, M. (2010). Contextualizing semantic representations using syntactically enriched vector models. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 948–957, Uppsala, Sweden, July. Association for Computational Linguistics.

Xiang, R., Chersoni, E., Lu, Q., Huang, C.-R., Li, W., and Long, Y. (2021). Lexical data augmentation for sentiment analysis. Journal of the Association for Information Science and Technology, 72, jun.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2020). Xlnet: Generalized autoregressive pretraining for language understanding.

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