Word Sense Disambiguation for Automatic Translation of Medical Dialogues into Pictographs

Magali Norré[1,2], Rémi Cardon[1], Vincent Vandeghinste[3], Thomas François[1]
[1]CENTAL, UCLouvain, Belgium
[2]FTI/TIM, Université de Genève, Switzerland
[3]Instituut voor de Nederlandse Taal, The Netherlands
Centre for Computational Linguistics, Leuven.AI, KU Leuven, Belgium
{magali.norre,remi.cardon,thomas.francois}@uclouvain.be
vincent.vandeghinste@ivdnt.org

Abstract

Word sense disambiguation is an NLP task embedded in different applications. We propose to evaluate its contribution to the automatic translation of French texts into pictographs, in the context of communication between doctors and patients with an intellectual disability. Different general and/or medical language models (Word2Vec, fastText, CamemBERT, FlauBERT, DrBERT, and CamemBERT-bio) are tested in order to choose semantically correct pictographs leveraging the synsets in the French WordNets (WOLF and WoNeF). The results of our automatic evaluations show that our method based on Word2Vec and fastText significantly improves the precision of medical translations into pictographs. We also present an evaluation corpus adapted to this task.

1 Introduction

Dialogue between doctors and patients is essential, as it enhances the patients' health status, their medication adherence, and their overall quality of life (Riedl and Schüßler, 2017). However, this dialogue can be impaired by misunderstandings, in particular for patients with an Intellectual Disability (ID). Various Augmentative and Alternative Communication (AAC) systems are used by people with disabilities (Beukelman and Mirenda, 1998), including automatic translation tools from text into pictographs (Vandeghinste et al., 2015).

One of the main issues that those systems face is polysemy. For example, in the French sentence to be translated *"avez-vous appliqué une crème sur la lésion ?"* (did you put cream on the lesion?), *"crème"* (cream) can be interpreted as OINTMENT or LIQUID CREAM. A translation system has to be able to produce the correct pictograph, here one that would represent OINTMENT.

In this article, we focus on Word Sense Disambiguation (WSD) of French polysemous words that can be used orally by doctors in questions and instructions for anamnesis in emergency settings (Norré et al., 2022). The Text-to-Picto system we use translates French into Arasaac,[1] Sclera[2] or Beta[3] pictograph sets, designed for AAC users with an ID (Norré et al., 2021). In order to provide a better semantic understanding of the input sentence, we test various language models (static, contextual, trained on general and/or medical data), and different French sense inventories. In addition, we present an evaluation corpus adapted to this task.

Section 2 describes existing work on WSD and text-to-pictograph systems. Section 3 introduces our methodology and the language models we used, while section 4 presents the Text-to-Picto system, the evaluation corpus, and the results. Our evaluations with Word2Vec and fastText show significant improvements over the baseline with the Text-to-Picto tool. We discuss the results in section 5.

2 Related Work

WSD has already been used in automatic text-to-pictograph systems, in order to improve the translation of polysemous words for the general language. For English, Mihalcea and Leong (2008) describe a basic WSD tool based on WordNet (Miller, 1995), but they do not evaluate its effectiveness within their text-to-pictograph translation system. Imam et al. (2019) test different WSD techniques – original Lesk, adapted Lesk, max similarity, Support Vector Machine (SVM) – with the English WordNet. They show that the system with the SVM obtains the best results (using recall, precision, and F-score). In Text-to-Picto, a system originally designed for Dutch (Vandeghinste et al., 2015), Sevens et al. (2016) use an external WSD tool,

[1]https://arasaac.org
[2]https://www.sclera.be
[3]https://www.betasymbols.com

based on SVM and developed within the framework of the DutchSemCor project (Vossen et al., 2012). Sevens (2018) specifically evaluated the contribution of this WSD tool using a corpus of 50 sentences that contain at least one ambiguous word. She obtained an improvement in precision for Sclera pictographs (from 29/50 to 41/50), and for Beta (from 28/50 to 42/50), demonstrating the added value of integrating a step of WSD.

For French, Vaschalde et al. (2018); Macaire et al. (2022) were the first to underline the importance of using WSD in a pictograph translation tool. Related to medical language, there are translation systems with pictographs, but they do not include WSD. This is the case of the French Text-to-Picto (Norré et al., 2022), but also for PictoDr, based on a neural translation approach using concepts, instead of words (Mutal et al., 2022; Gerlach et al., 2023). We therefore aim to assess the contribution of WSD in the context of specialized language for automatic translation into pictographs, an issue that has not yet been addressed in the literature.

3 Methodology

We present our WSD algorithm below, through the example of the noun *"alcool"* (alcohol) to be disambiguated (Figure 1) in the sentence *"avez-vous bu de l'alcool ?"* (did you drink alcohol?). The two possible translations into an Arasaac pictograph are: ALCOHOLIC DRINK, and ISOPROPYL ALCOHOL. The lemma *"alcool"* refers to three different synsets in WOLF (Sagot and Fišer, 2008), the French WordNet used by default in the Text-to-Picto system (Norré et al., 2022).

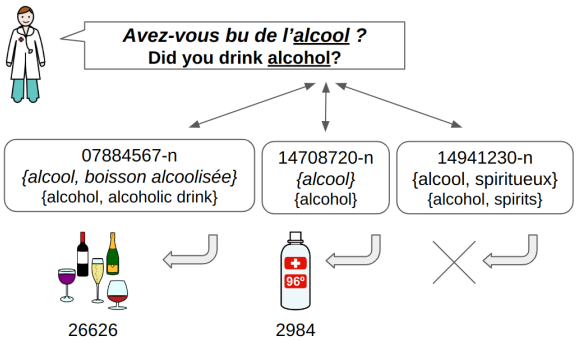

Figure 1: Pictographs for the word to be disambiguated: *"alcool"* (alcohol). Ids are indicated for the WOLF synsets and the Arasaac pictographs.

We differentiate steps using static embeddings and contextual embeddings by marking them respectively with (a) and (b).

1. (a) Retrieve in Word2Vec (Mikolov et al., 2013), or fastText (Bojanowski et al., 2017) the vectors of lemmas (content words) of the input sentence, i.e., nouns, verbs, adjectives, and adverbs – tagged with TreeTagger (Schmid, 1994). We average these vectors in order to get a contextual representation from a static representation (sentence vector).

   (b) Retrieve in CamemBERT (Martin et al., 2020), FlauBERT (Le et al., 2020), DrBERT (Labrak et al., 2023), or CamemBERT-bio (Touchent et al., 2023)[4] a vector of lemmas (content words) of the input sentence in order to use them as context (sentence vector).

2. (a) For each synset i (from 1 to N) linked to the polysemous lemma in the French WordNet, retrieve all lemmas having the following semantic relations – synonyms, hyperonyms, hyponyms, and near synonyms with a different part-of-speech tag (eng_derivative relation) – with the lemma. Then, get the distributed representations of all these semantically related words in Word2Vec or fastText and average them to get a contextual static representation of each synset i (relation vector).

   (b) Similarly, for each synset i (from 1 to N), get the list of semantically related words as in 2a, and join them as a unique string. Then, retrieve in CamemBERT, FlauBERT, DrBERT, or CamemBERT-bio a contextual vector representing each synset i (relation vector).

3. Calculate the cosine similarities between the sentence vector and the relation vector of each synset i.

   Example: {'synset1' (07884567-n): 0.64, 'synset2' (14708720-n): 0.35, 'synset3' (14941230-n): 0.25}

4. Use the cosine scores to select the pictograph(s) to retrieve. We rank the synsets, sorted by cosine similarity in descending order. We start by retrieving the pictograph(s) of the synset that comes first (rank 1), if

[4] In French. An English version is available here: https://arxiv.org/abs/2306.15550.

this synset is not linked to a pictograph, we retrieve the pictograph(s) of the synset that comes after, and so on until a pictograph is found (rank > 1).

Example: {'synset1' (07884567-n): 26626, 'synset2' (14708720-n): 2984, 'synset3' (14941230-n): -}

For our language models, we used pre-trained models for French (Table 1): frWac2Vec and frWiki2Vec for Word2Vec (Fauconnier, 2016);[5] Common Crawl + Wikipedia for fastText (Grave et al., 2018).[6] frWac2Vec is a collection of embeddings trained on the frWaC corpus (Baroni et al., 2009), which is composed of 1.6 billion words. It was built from the web. The crawl was limited to the .fr domain, while using medium frequency words from the *Le Monde Diplomatique* corpus and basic French vocabulary lists.[7] The frWiki2Vec corpus was trained on 600 million words. frWac2Vec is available in 12 different versions (lemmatized or not, part-of-speech tagged or not, CBOW or Skip-Gram, with vectors of different dimensions and various minimum frequencies of words in the corpora). There are also 8 versions of frWiki2Vec. We used the 500-dimension models, lemmatized with TreeTagger, but not tagged. We tested all the pre-trained models of CamemBERT,[8] FlauBERT,[9] DrBERT[10] and CamemBERT-bio.[11] The DrBERT models are specific to the medical domain, as they were trained on the NACHOS corpus (Labrak et al., 2023), which consists of 24 biomedical resources under free license. This is also the case for CamemBERT-bio, a state-of-the-art language model trained on a French public biomedical corpus (Touchent et al., 2023). It was built using continual-pretraining from CamemBERT.

We also trained 500-dimension Word2Vec and fastText models – CBOW and Skip-Gram –, on the CLEAR corpus (Grabar and Cardon, 2018),[12] using the same Word2Vec hyperparameters as Cardon (2021, p. 47).[13] For training with fastText, we used the default hyperparameters. CLEAR is a French medical corpus consisting of three sub-corpora: articles from online encyclopedias (Wikipedia and Vikidia), drug leaflets, and summaries of the Cochrane Foundation's medical scientific literature. It is a comparable corpus, with texts in a technical version and in a simple/simplified version. We used the three sub-corpora, once with medical encyclopedia articles (146 million words in total) and another time adding general articles (+65 million words). We did not pre-process this corpus before training. Note that CLEAR is a part of the NACHOS and biomed-fr corpora that were used to train DrBERT and CamemBERT-bio.

We therefore propose to evaluate several language models, trained with general and/or medical data (Table 1). We compare Word2Vec and fastText to contextual BERT models for French. We also test two French WordNets: WOLF and WoNeF (Pradet et al., 2014).

4 Evaluation

In this section we describe our baseline – i.e., the pictograph translation tool without WSD – (section 4.1), our evaluation corpus (section 4.2), and the results (section 4.3).

4.1 Pictograph Translation System

In order to evaluate our hypothesis, i.e., WSD improves the precision of pictograph translation, we used the Text-to-Picto system (Vandeghinste et al., 2015; Sevens, 2018), adapted to French (Norré et al., 2021, 2022). In this tool, the source text first undergoes a shallow linguistic analysis (Figure 2): it is tokenized, part-of-speech tagged, and lemmatized with TreeTagger.

Two routes are possible to translate text into pictographs: the direct route and the semantic route. In the direct route, the lemma is looked up in a pictograph dictionary and directly translated into a pictograph. In the semantic route, French WordNet is used as a pivot: synsets related to the lemma are identified and connected to pictographs. More precisely, if the word is a noun, verb, adjective or adverb, it is looked up in WOLF. We also use WordNet relations – such as hyperonyms, hyponyms, antonyms, and near synonyms with a different part-of-speech tag – to retrieve semantically-related

[5] https://fauconnier.github.io

[6] https://fasttext.cc/docs/en/crawl-vectors

[7] https://wacky.sslmit.unibo.it/doku.php?id=corpora

[8] https://camembert-model.fr

[9] https://github.com/getalp/Flaubert

[10] https://drbert.univ-avignon.fr

[11] https://huggingface.co/almanach/camembert-bio-base

[12] http://natalia.grabar.free.fr/resources.php#clear

[13] -window 7 -sample 1e-5 -hs 1 -negative 50 -mincount 20 -alpha 0.025.

#	Model	Corpus	(#) Param.	Dim.	# Types | GB
1a	Word2Vec	frWac2Vec	CBOW	500	119,227
1b	Word2Vec	frWac2Vec	Skip	500	119,227
2a	Word2Vec	frWiki2Vec	CBOW	500	66,819
3a	Word2Vec	CLEAR (medical + general)	CBOW	500	198,164
3b	Word2Vec	CLEAR (medical + general)	Skip	500	198,164
4a	Word2Vec	CLEAR (medical)	CBOW	500	79,456
4b	Word2Vec	CLEAR (medical)	Skip	500	79,456
5a	fastText	Common Crawl + Wikipedia	CBOW	300	?
6a	fastText	CLEAR (medical + general)	CBOW	500	198,164
6b	fastText	CLEAR (medical + general)	Skip	500	198,164
7a	fastText	CLEAR (medical)	CBOW	500	79,456
7b	fastText	CLEAR (medical)	Skip	500	79,456
8A	CamemBERT (base)	OSCAR	110 M	768	138 GB
8B	CamemBERT (base)	OSCAR (sample)	110 M	768	4 GB
8C	CamemBERT (base)	CCNet	110 M	768	135 GB
8D	CamemBERT (base)	CCNet (sample)	110 M	768	4 GB
8E	CamemBERT (base)	Wikipedia	110 M	768	4 GB
8F	CamemBERT (large)	CCNet	335 M	1,024	135 GB
9A	FlauBERT (base, uncased)	Diverse (Wikipedia, books, etc.)	137 M	768	71 GB
9B	FlauBERT (base, cased)	Diverse (Wikipedia, books, etc.)	138 M	768	71 GB
9C	FlauBERT (large, cased)	Diverse (Wikipedia, books, etc.)	373 M	1,024	71 GB
9D	FlauBERT (small, cased)	Diverse (Wikipedia, books, etc.)	54 M	512	71 GB
10A	DrBERT (base, cased)	NACHOS (large)	110 M	768	7.4 GB
10B	DrBERT (base, cased)	NACHOS (small)	110 M	768	4 GB
10C	DrBERT (base, cased)	NACHOS (small-PubMedBERT)	110 M	768	4 GB
10D	DrBERT (base, cased)	NACHOS (small-CamemBERT)	110 M	768	4 GB
11A	CamemBERT-bio (base)	biomed-fr	110 M	768	2.7 GB

Table 1: Language models: Word2Vec, fastText, CamemBERT, FlauBERT, DrBERT, and CamemBERT-bio.

synsets. Based on the synsets selected, pictographs are generated using the database of Norré et al. (2021). To choose the optimal path while converting a sequence of lemmas to a sequence of pictographs, a search algorithm A* is used, described in detail by Vandeghinste et al. (2015). It works with different parameters (i.e., penalties) related to WordNet relations, pictograph features, and route preference. When pictographs have the same weight at the end, they are sorted according to their names and the first is chosen.

We are looking for a way to improve the semantic route that would also replace the search algorithm of this translation system and rank synsets based on the context of the input text. We focus here on polysemous words, the others (e.g. the pronoun in Figure 1) being likely to be translated into a pictograph with the direct route of the tool.

4.2 Evaluation Corpus

To build an evaluation corpus adapted to our task, we automatically translate several hundred French sentences from the BabelDr medical speech translation system (Bouillon et al., 2021) with Text-to-Picto. We use the AZ (pictograph names sorted in alphabetical order) and ZA (reverse) modes. We do so in order to detect words with at least two possible translations in Arasaac belonging to the same grammatical category as the ambiguous word. We sample 100 polysemous lemmas,[14] and extract, for each of them, at least one sentence from the BabelDr system – containing at least two lemmas which are a NOUN, VER, ADJ or ADV (the average number of lemmas per sentence is 3.67) –, at least one Arasaac pictograph with a correct sense, one Arasaac pictograph with an incorrect sense and their WOLF synsets.

We deliberately avoided multi-word expressions that are used as pictograph names by Arasaac, because we believe that a specific linguistic processing in order to automatically translate them by a single pictograph would be required. This is the case of *"prise de sang"* (blood test) incorrectly translated by two pictographs (Norré et al., 2022, pp. 47-48): *"tenir"* (grasp) + *"sang"* (blood). Those expressions can generate ambiguity problems in the Text-to-Picto system if they are not

[14] Our evaluation is based on that of Sevens (2018), i.e., the test point method (Shiwen, 1993). "A test point is a specific problem which an MT system has to resolve. In the test point method, for each test sentence, substring matching is used to determine if the specific test point has been correctly processed" (Sevens, 2018, p. 164).

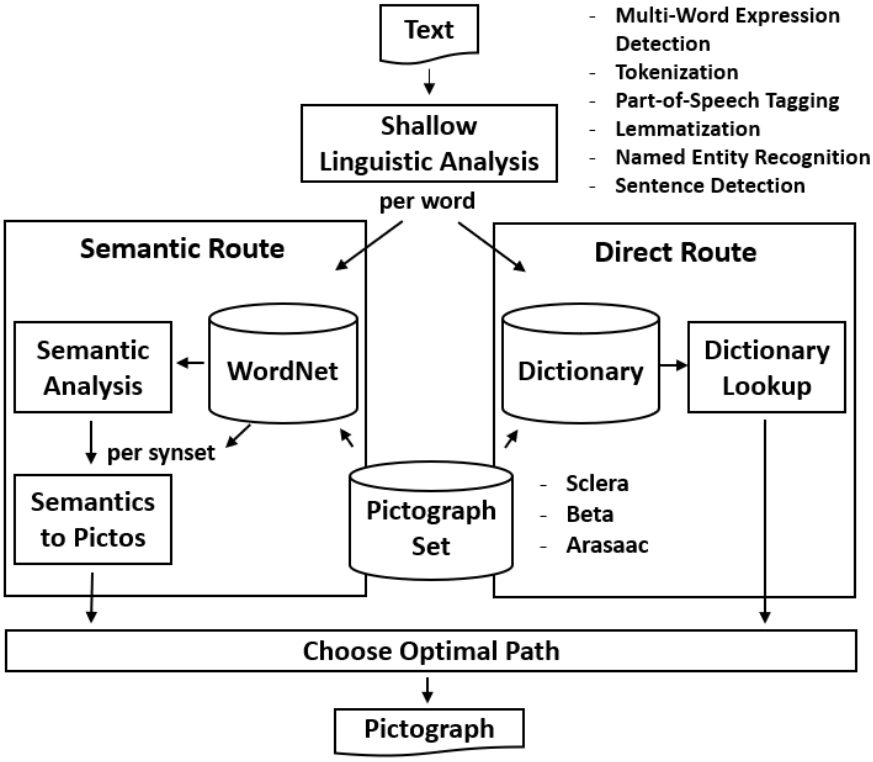

Figure 2: Architecture of the French Text-to-Picto tool (Norré et al., 2021), adapted from Vandeghinste et al. (2015).

specifically encoded in a dictionary or annotated with two WordNet synsets.

On average, the 100 polysemous words in our corpus are linked to 13.49 synsets. The minimum is 2 synsets (for the noun *"seringue"*, syringe), and the maximum is 102 (for the verb *"donner"*, give, versus 44 for "give" in the Open English WordNet).[15] Our evaluation corpus consists of 52 nouns, 38 verbs, 5 adjectives, and 5 adverbs.

	Precision	Recall	F1
Arasaac			
AZ	0.35	0.99	0.52
ZA	0.45	0.99	0.62
Average	0.40	0.99	0.57

Table 2: Precision, Recall, and F1 scores of Text-to-Picto without WSD (in AZ and ZA modes) on 100 polysemous words for Arasaac pictographs with WOLF.

4.3 Results

First, we evaluated the precision of Arasaac translations for our 100 polysemous words, generated in AZ and ZA modes by the Text-to-Picto system without a WSD module [16] (Table 2). Precision varies between 0.35 and 0.45 depending on the sort method. Recall is the percentage of translated words. F1 scores vary between 0.52 and 0.62.

Then, we automatically computed recall by limiting ourselves to the pictograph(s) of the synset with rank 1 (see section 3). However, it should be noted that many of these synsets are not linked to an Arasaac, Sclera or Beta pictograph (Figure 3).

The rank 1 method yields a low recall (in range 0.32–0.50, depending on the language model, for Arasaac, and in range 0.15–0.29 for Sclera or Beta). We observe that the rank > 1 method yields the same recall as the Text-to-Picto system without WSD: around 1.0 for Arasaac (Figure 3). Sclera and Beta have a recall between 0.73–0.76. This underlines the importance of being able to look for more than one acceptable synset, to account for the rather low coverage of the pictograph sets.

We also automatically evaluated the precision of all our WSD models based on the correct synsets of each polysemous word translated into an Arasaac pictograph (Table 3). To do so we compared each synset obtained against the evaluation corpus. Pictographs – from the same set – linked to different synsets were sometimes accepted for the same word, because they were adapted to the context of

[15] https://en-word.net/lemma/give

[16] With the following optimized parameters: -penal 9 -hyper 15 -anto 10 -oov 3 -dict 2.

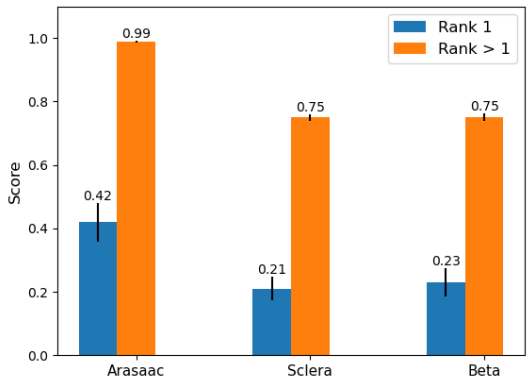

Figure 3: Recall scores of WSD (rank 1 and rank > 1) on 100 polysemous words for Arasaac, Sclera, and Beta pictographs with WOLF.

the sentence as in example (a) in Figure 4, for the sentence *"avez-vous d'autres problèmes de santé ?"* (do you have any other health problems?). As a baseline, we use Text-to-Picto without WSD in the AZ mode (the default mode in Text-to-Picto) on the same 100 words, for Arasaac (see Table 2).

#	P	Rel. improv.	#	P	Rel. improv.
Baseline	**0.35**	–	8A	0.45	+0.10
1a	0.66	+0.31**	8B	0.41	+0.06
1b	**0.73**	**+0.38****	8C	**0.48**	**+0.13**
2a	0.53	+0.18**	8D	0.41	+0.06
3a	0.53	+0.18*	8E	0.46	+0.11
3b	0.58	+0.23**	8F	0.44	+0.09
4a	0.62	+0.27**	9A	0.44	+0.09
4b	0.61	+0.26**	9B	0.45	+0.10
5a	**0.66**	**+0.31****	9C	0.39	+0.04
6a	0.56	+0.21**	9D	**0.49**	**+0.14***
6b	0.65	+0.30**	10A	**0.45**	**+0.10**
7a	0.60	+0.25**	10B	0.42	+0.07
7b	0.63	+0.28**	10C	**0.45**	**+0.10**
			10D	0.42	+0.07
			11A	**0.46**	**+0.11**

* $p < 0.05$, ** $p < 0.01$

Table 3: Precision scores and Relative improvement of WSD (rank > 1) on 100 polysemous words for Arasaac pictographs with WOLF. References for language models can be found in Table 1.

The model that obtains the best precision is the Word2Vec Skip-Gram with the frWac2Vec corpus (1b: 0.73), followed by the the same model in CBOW version (1a: 0.66), as well as the fastText for general language (5a: 0.66), then the fastText Skip-Gram model that we trained on the medical and general part of the CLEAR corpus (6b: 0.65). It is important to note that our best model (1b) obtains a precision of 0.73, a relative improvement of +0.38 over the performance of the actual French Text-to-Picto system for Arasaac without WSD. Note that Baseline, 4a, and 4b have a recall of 0.99.

To show the contribution of WSD, we present examples of problematic pictographs with the Text-to-Picto system in AZ or ZA modes for 4 words to be disambiguated (Figure 4). The pictograph on the left represents the correct sense, the one on the right the incorrect sense. The most appropriate pictograph for the adjective *"autre"* (other, ex. a) in our sentence was linked to two synsets. We have therefore accepted both of them (02069355-a and 02070188-a). With our WSD methods, the case (a) was still wrongly translated by the pictograph *"nouveau"* (new, ex. b) in our 27 models. However, the correct sense of the noun *"cœur"* (heart, ex. c), the verb *"opérer"* (operate, ex. e),[17] and the adverb *"souvent"* (often, ex. g) was selected in 16, 18, and 23 models, respectively.

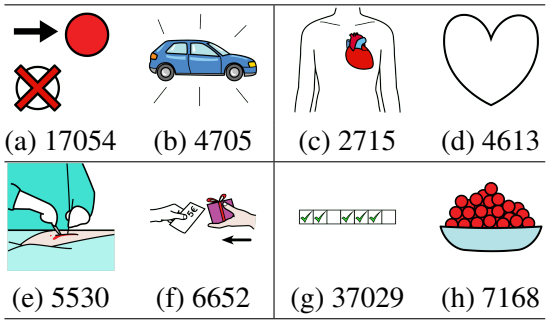

Figure 4: Arasaac pictographs: example of words to be disambiguated (a-b) *"autre"* (other), (c-d) *"cœur"* (heart), (e-f) *"opérer"* (operate), (g-h) *"souvent"* (often)

We evaluated these models on WOLF, but also on the three different versions of another French WordNet, WoNeF. WOLF and WoNeF are two automatic translations of the Princeton WordNet 3.0, they differ in the way they were built.[18] Our results confirm that they are very different, WOLF being better in recall, precision, and F1 (Figure 5). If we compare the WoNeFs with each other, on average, the high "coverage" version gets the best recall (0.5), the high "f-score" version has the best precision (0.43), while the F1 of small "precision" version is extremely limited (0.1).

[17] Linked to the synset {*opérer, vendre, commercialiser, distribuer, échanger*} ({operate, sell, market, distribute, exchange}), the pictograph *"vendre"* (sell, ex. f) – the bad translation – is selected because of the expression *"opérer une transaction"* (operate a transaction).

[18] As noted by Norré et al. (2021), the three versions of WoNeF are the result of optimizing the three metrics. The high coverage version contains 109,447 pairs (literal, synset), the main WoNeF has an F-score of 70.9%, and the high precision version has a precision of 93.3% (Pradet et al., 2014).

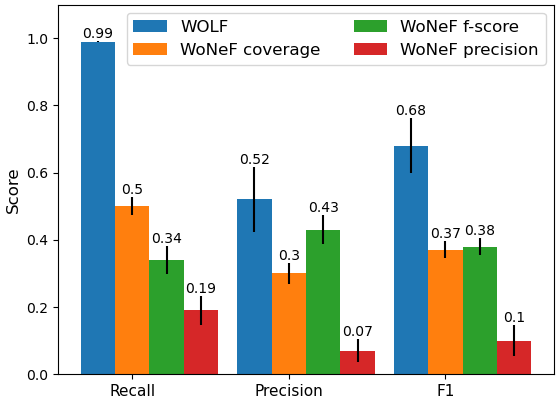

Figure 5: Recall, Precision, and F1 scores of WSD (rank > 1) on 100 polysemous words for Arasaac pictographs with WOLF and WoNeF.

5 Discussion

We evaluated the impact of different training corpora (e.g. general language, medical language, etc.) on our performance. Models pre-trained on general language data obtain higher precision on average (0.64 for the frWac2Vec and frWiki2Vec models, and 0.66 for fastText trained on Common Crawl + Wikipedia) than CLEAR models (medical + general: 0.58; medical: 0.61). Beyond the effect of the size of the training corpus of these pre-trained models, another explanation for these counter-intuitive results may be the fact that even if the words to be disambiguated are integrated into medical dialogues, they are not all medical terms: out of our 100 polysemous words, only 19% are found in the medical Wiktionary extracted by Cardon (2018), 56% in the medical lexicon of Grabar and Hamon (2016), and 29% in the SNOMED International terminology (Côté, 1996); 37% of them in our corpus are in at least two of these three resources.

Regarding the performance of the language models, we found that the average precision of the five fastText models (0.62) is very close to that of the seven Word2Vec models (0.61). The lower averages of CamemBERT (0.44), FlauBERT (0.44), DrBERT (0.43), and CamemBERT-bio (0.46) are counter-intuitive and we hypothesized that the small size of our input context could be a factor. To verify this, we performed experiments leveraging context for disambiguating the lemmas. We extracted the usages (<USAGE>) in WOLF (# 48,233), i.e., syntagms or short sentences that serve as examples of use. As they are only available in English (directly transferred from WordNet), we automatically translated into French the 649 usages associated to the lemmas in our evaluation corpus, with Google Translate. For example, the usage of WOLF translated from English (alcohol (or drink) ruined him) into French is *"l'alcool (ou la boisson) l'a ruiné"* for the word *"alcool"* (see Figure 1). We tested several encoding configurations for the 15 BERT language models with WOLF (Table 4).

There were two configurations for the sentence vector (step 1b): A) lemmas of the content words (e.g., *"avoir boire alcool"*); B) the whole sentence (*"avez-vous bu de l'alcool ?"*). For the relation vector (step 2b), we tested six configurations: a) words of the 4 types of relations; b) words of the 4 types of relations, each followed by a period; c) the usages; d) the usages followed by a period; e) the usages followed by a period and synonyms,[19] each followed by a period; f) the usages followed by a period and words of the 4 types of relations, each followed by a period.

Depending on these encoding configurations, the precision of our models can vary from -0.14 to +0.20 compared to our main method, i.e. our BERT results with parameters A-a (see Table 3).[20] Using only the usages (A/B-c°/d°), we obtained a recall of 0.66 on our 100 words. The configuration with a recall of 1.0 and the highest average precision is the B-e (with 0.47 vs. 0.44 for A-a). Even if we observe improvements compared to the main method, the BERT language models remain less precise than Word2Vec and fastText.

The sentences can be useful for BERT contextual models to improve the precision (A/B-c°/d°). We have however noted that encoding only usages as relation vectors is not efficient, because not enough of them are associated with synsets linked to Arasaac pictographs (recall: 0.66). Therefore, usages must be combined with synonyms (B-e). BERT language models applied to the WOLF data with our method, however, do not offer a great improvement in precision if we compare them to the Text-to-Picto system in ZA mode, which obtains 0.45 on the same 100 polysemous words (see Table 2). Another room for improvement would be to use the French SemCor (Nasiruddin et al., 2015), but these data are not adapted to medical dialogue.

[19] Encoding a sentence followed by a list of words as BERT input is a technique that shows promising results for lexical simplification (Wilkens et al., 2022).

[20] From 0.48 to 0.34 for 8C (B-a), and from 0.45 to 0.65 for 10A (A-c°).

	Precision															
Parameter	**8A**	**8B**	**8C**	**8D**	**8E**	**8F**	**9A**	**9B**	**9C**	**9D**	**10A**	**10B**	**10C**	**10D**	**11A**	**Avg.**
A-a	0.45	0.41	0.48	0.41	0.46	0.44	0.44	0.45	0.39	**0.49***	0.45	0.42	0.45	0.42	0.46	0.44
A-b	0.41	0.40	0.44	0.39	0.38	0.41	0.36	0.45	0.45	**0.49**	**0.49***	0.42	0.41	0.39	0.41	0.42
A-c°	0.51	0.53	0.50	0.59*	0.53	0.57*	0.53	0.48	0.59*	0.56*	**0.65****	0.56*	0.57*	0.46	0.48	°0.54
A-d°	0.57*	0.50	0.50	0.53	0.53	**0.60****	0.53	0.51	0.51	0.56*	0.59*	0.51	0.56*	0.46	0.48	°0.53
A-e	0.43	0.46	**0.48**	0.45	0.45	0.44	0.42	0.43	0.39	**0.48**	0.42	0.39	0.41	0.36	0.38	0.42
A-f	0.44	0.43	**0.48**	0.42	0.35	0.46	0.41	0.43	0.44	0.45	0.40	0.43	0.39	0.38	0.41	0.42
B-a	0.44	0.38	0.34	0.42	0.43	0.49*	0.43	0.39	0.33	**0.50***	0.36	0.36	0.42	0.33	0.43	0.40
B-b	0.41	0.41	0.43	0.43	0.36	0.45	0.46	0.42	0.38	**0.47**	0.32	0.38	0.41	0.43	0.42	0.41
B-c°	0.48	**0.57***	0.54*	0.51	0.56*	0.50	0.54	0.46	0.53	0.48	0.54*	0.46	0.51	0.45	0.51	°0.51
B-d°	0.51	0.51	0.53	0.53	0.50	0.54	0.53	**0.59***	0.56*	0.53*	0.50	0.53	0.53	0.50	0.50	°0.52
B-e	0.44	0.48	**0.53****	0.51*	0.48	0.50*	0.41	0.46	0.47	0.47	**0.53****	0.43	0.45	0.42	0.45	0.47
B-f	0.44	0.43	**0.48**	0.42	0.35	0.46	0.41	0.43	0.44	0.45	0.40	0.43	0.39	0.38	0.37	0.42
Avg. by model	0.46	0.45	0.47	0.46	0.44	0.48	0.45	0.45	0.45	0.49	0.47	0.44	0.45	0.41	0.44	
Avg. by family	0.46						0.46				0.44				0.44	

* $p < 0.05$, ** $p < 0.01$

Table 4: Precision scores of WSD (rank > 1) on 100 polysemous words for Arasaac pictographs with WOLF, BERT models, and various encoding parameters. References for language models can be found in Table 1.

Finally, we compared several French WordNets (see Figure 5). Each of them produced rather different pictographs, due to different synset scopes. Norré et al. (2021) already showed that better results can be reached with WOLF than with the three versions of WoNeF using the Text-to-Picto system for the Arasaac pictograph set. In WOLF and two versions of WoNeF ("coverage" and "f-score"), only half of the English WordNet synsets have been translated into French.

Choosing an appropriate synset is not always enough to get a correct translation. It would also be necessary to refine the selection of pictographs within the synset obtained with WSD. This is the case of Arasaac where many pictographs – sometimes twenty – can be associated with a single synset. They can be identical pictographs (with a character who is non-gendered, male or female), but also with a more or less different meaning although they belong to the same synset (e.g. "lift" the toilet seat, a baby, an object, etc.). We do not have information about the method used by Arasaac to label the pictographs.

6 Conclusion

In this paper, we performed experiments for WSD with different language models, either static (Word2Vec, fastText), or contextual (CamemBERT, FlauBERT, DrBERT, CamemBERT-bio), in medical French. We observed that the most promising method is to use Word2Vec or fastText in order to improve the precision of translations into pictographs (see Table 3). According to our experiments, the effectiveness of contextual language models is rather limited compared to static vector representations for this task. The advantage of our method is that it is easily applicable to other natural languages that have medium-sized corpora – which can be used to train Word2Vec or fastText – and a WordNet. We have also built and made available the first evaluation corpus for the WSD of medical sentences into Arasaac pictographs.[21]

There is room for further improvement to adapt our approach. For example, we could test other operations than the average in order to produce a contextual representation from static vectors. It would also be possible to use other relations in WOLF, beyond synonyms, hyperonyms, hyponyms, and near synonyms. WOLF and the three WoNeFs offer 18 exploitable relations. Finally, another perspective to improve the system would be to perform WSD based on the filenames or other metadata of the pictographs and the French resource of disambiguated synonyms, ReSyf (François et al., 2016).

Acknowledgments

Magali Norré is supported by the Université catholique de Louvain (UCLouvain) FSR mandate N°17005.2022. Rémi Cardon is supported by the FSR Incoming Postdoc Fellowship program of UCLouvain. This work is also part of the PROPICTO project, funded by the Fonds National Suisse (N°197864) and the Agence Nationale de la Recherche (ANR-20-CE93-0005). The pictographs used are property of the Aragon Government and have been created by Sergio Palao to Arasaac. Aragon Government distributes them under Creative Commons License.

[21]The evaluation corpus and source code are available for the research community at the following address: https://github.com/VincentCCL/Picto.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The Wacky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

David R. Beukelman and Pat Mirenda. 1998. *Augmentative and alternative communication*. Paul H. Brookes Baltimore.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Pierrette Bouillon, Johanna Gerlach, Jonathan David Mutal, Nikolaos Tsourakis, and Hervé Spechbach. 2021. A Speech-enabled Fixed-phrase Translator for Healthcare Accessibility. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 135–142. Association for Computational Linguistics.

Rémi Cardon. 2018. Approche lexicale de la simplification automatique de textes médicaux. In *Actes de la Conférence TALN*, pages 159–174.

Rémi Cardon. 2021. *Simplification automatique de textes techniques et spécialisés*. Ph.D. thesis, Université de Lille.

Roger A. Côté. 1996. Répertoire d'anatomopathologie de la SNOMED internationale, v3. 4. *Université de Sherbrooke, Sherbrooke, Québec*.

Jean-Philippe Fauconnier. 2016. *Acquisition de liens sémantiques à partir d'éléments de mise en forme des textes*. Ph.D. thesis, Université de Toulouse.

Thomas François, Mokhtar Billami, Núria Gala, and Delphine Bernhard. 2016. Bleu, contusion, ecchymose : tri automatique de synonymes en fonction de leur difficulté de lecture et compréhension. In *JEP-TALN-RECITAL 2016*, volume 2, pages 15–28.

Johanna Gerlach, Pierrette Bouillon, Magali Norré, and Hervé Spechbach. 2023. Translating Medical Dialogues into Pictographs: An Approach Using UMLS. In *Caring is Sharing – Exploiting the Value in Data for Health and Innovation. Proceedings of the 33rd Medical Informatics Europe Conference*, pages 823–824, Gothenburg, Sweden. European Federation for Medical Informatics (EFMI) and IOS Press.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – Simple Corpus for Medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation*, pages 3–9. Association for Computational Linguistics.

Natalia Grabar and Thierry Hamon. 2016. A Large Rated Lexicon with French Medical Words. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2643–2648, Portorož, Slovenia. European Language Resources Association.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mai Farag Imam, Amal Elsayed Aboutabl, and Ensaf H. Mohamed. 2019. Automating Text Simplification Using Pictographs for People with Language Deficits. *I.J. Information Technology and Computer Science*, 9(1):26–34.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Cécile Macaire, Lucía Ormaechea Grijalba, and Adrien Pupier. 2022. Une chaîne de traitements pour la simplification automatique de la parole et sa traduction automatique vers des pictogrammes. In *Actes de la 29e conférence sur le Traitement Automatique des Langues Naturelles*, pages 111–123. ATALA.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Eric De la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219. Association for Computational Linguistics.

Rada Mihalcea and Chee Wee Leong. 2008. Toward Communicating Simple Sentences Using Pictorial Representations. *Machine Translation*, 22(3):153–173.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 26. Curran Associates Inc.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Jonathan David Mutal, Pierrette Bouillon, Magali Norré, Johanna Gerlach, and Lucía Ormaechea Grijalba. 2022. A Neural Machine Translation Approach to

Translate Text to Pictographs in a Medical Speech Translation System – The BabelDr Use Case. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas*, pages 252–263, Orlando, USA. Association for Machine Translation in the Americas.

Mohammad Nasiruddin, Andon Tchechmedjiev, Hervé Blanchon, and Didier Schwab. 2015. Création rapide et efficace d'un système de désambiguïsation lexicale pour une langue peu dotée. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, pages 83–94.

Magali Norré, Vincent Vandeghinste, Pierrette Bouillon, and Thomas François. 2021. Extending a Text-to-Pictograph System to French and to Arasaac. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1050–1059.

Magali Norré, Vincent Vandeghinste, Thomas François, and Pierrette Bouillon. 2022. Investigating the Medical Coverage of a Translation System into Pictographs for Patients with an Intellectual Disability. In *Proceedings of SLPAT 2022: 9th Workshop on Speech and Language Processing for Assistive Technologies*, pages 44–49. Association for Computational Linguistics.

Quentin Pradet, Gaël De Chalendar, and Jeanne Baguenier Desormeaux. 2014. WoNeF, an improved, expanded and evaluated automatic French translation of WordNet. In *Proceedings of the Seventh Global Wordnet Conference*, pages 32–39.

David Riedl and Gerhard Schüßler. 2017. The Influence of Doctor-Patient Communication on Health Outcomes: A Systematic Review. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, 63(2):131–150.

Benoît Sagot and Darja Fišer. 2008. Building a free French WordNet from multilingual resources. In *OntoLex*, Marrakech, Morocco.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Leen Sevens. 2018. *Words Divide, Pictographs Unite: Pictograph Communication Technologies for People with an Intellectual Disability*. LOT, Utrecht, The Netherlands.

Leen Sevens, Gilles Jacobs, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde. 2016. Improving Text-To-Pictograph Translation Through Word Sense Disambiguation. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 131–135. Association for Computational Linguistics.

Yu Shiwen. 1993. Automatic Evaluation of Output Quality for Machine Translation Systems. *Machine Translation*, 8(1):117–126.

Rian Touchent, Laurent Romary, and Eric De La Clergerie. 2023. CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. In *30e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 323–334, Paris, France. ATALA.

Vincent Vandeghinste, Ineke Schuurman, Leen Sevens, and Frank Van Eynde. 2015. Translating text into pictographs. *Natural Language Engineering*, 23(2):217–244.

Céline Vaschalde, Pauline Trial, Emmanuelle Esperança-Rodier, Didier Schwab, and Benjamin Lecouteux. 2018. Automatic pictogram generation from speech to help the implementation of a mediated communication. In *Proceedings of the 2nd Swiss Conference on Barrier-free Communication*.

Piek Vossen, Attila Görög, Rubén Izquierdo, and Antal van den Bosch. 2012. DutchSemCor: Targeting the ideal sense-tagged corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 584–589, Istanbul, Turkey. European Language Resources Association.

Rodrigo Wilkens, David Alfter, Rémi Cardon, Isabelle Gribomont, Adrien Bibal, Patrick Watrin, Marie-Catherine de Marneffe, and Thomas François. 2022. CENTAL at TSAR-2022 Shared Task: How Does Context Impact BERT-Generated Substitutions for Lexical Simplification? In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 231–238.