Predicting pronouns across languages with continuous word spaces

Ngoc-Quan Pham
Erasmus Mundus Master Program in
Language and Communication Technology
ngoc-quan.pham.14@um.edu.mt

Lonneke van der Plas
Institute of Linguistics
University of Malta
lonneke.vanderplas@um.edu.mt

Abstract

Predicting pronouns across languages from a language with less variation to one with much more is a hard task that requires many different types of information, such as morpho-syntactic information as well as lexical semantics and coreference. We assumed that continuous word spaces fed into a multi-layer perceptron enriched with morphological tags and coreference resolution would be able to capture many of the linguistic regularities we found. Our results show that the model captures most of the linguistic generalisations. Its macro-averaged F-score is among the top-3 systems submitted to the DiscoMT shared task reaching 56.5%.

1 Introduction

This paper provides the description for the classification system, submitted by the University of Malta, to the DiscoMT shared task on cross-lingual pronoun prediction (Hardmeier et al., 2015). In this task, we are concerned with finding the correct French translations for the English third-person subject pronouns *it* and *they*. An example would be the following, where we need to predict the pronoun corresponding to the placeholder "REPLACE" given in the French sentence.

– And so, if you depend on these sources, you have to have some way of getting the energy during those time periods that **it**'s not available .
– Et donc, si vous dépendez de ces sources, vous devez avoir un moyen d'obtenir de l'énergie pendant ces périodes de temps où **REPLACE** n'est pas disponible.

The task is setup in such a way that the system needs to choose between 9 classes of French pronouns : *ce, elle, elles, il, ils, ça, cela, on*, and OTHER in bitexts in which the pronoun aligned to the English pronouns *it* and *they* are substituted by placeholders [1]. The difficulty of this task lies in the fact that the French translation for a particular English pronoun is generally inconsistent and dependent on many different factors. By analysing the linguistic characteristics of this problem, we identified the factors contributing to the predictability of the pronouns, as described in Section 2.

The dependencies are modeled by using a probabilistic neural network, motivated by previous work in the field of Statistical Language Modeling and Statistic Machine Translation. Specifically, the feature words are treated through a projection layer to become continuous vectors. This approach leads to a *distributed representation* of the words, that has shown to capture morpho-syntactic and semantic information (Mikolov et al., 2013c; Köper et al., 2015). After that, the output of the network is a soft-max layer computing probabilities of the possible outputs, such as language models (Bengio et al., 2003), or translation models (Son et al., 2012). The input words can belong to one single language (language model case (Bengio et al., 2003)), or even two different languages (translation model case (Son et al., 2012)). More importantly, the size of projected vectors is much smaller than the vocabulary, aiming at a reduction of the data sparseness problem. We apply the concept in our system, by learning the probabilities of the pronouns given the word vectors in the input layer.

In the works mentioned to motivate this structure, this projection layer is learned together with the neural network parameters (Schwenk, 2007; Mikolov et al., 2010; Le et al., 2011). For the task of cross-lingual pronoun prediction, Hardmeier et al. (2013) also chose to learn the projection matrices and the neural network weights at the same

1. For more information on the task setup we refer to the introductory paper of the shared task (Hardmeier et al., 2015)

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 101–107,
Lisbon, Portugal, 17 September 2015. ©2015 Association for Computational Linguistics.

time. We chose to train the projection matrix separately, and then train the neural network on top of the learned continuous word vectors (Mikolov et al., 2013a) to alleviate the training process.

In contrast to English, the source language in this shared task, every noun in French has a grammatical gender. Pronouns agree in gender and number with their antecedents (or postcedents). As a consequence, in many cases in which the English translation contains the pronoun *it*, we need to choose between *elle* or *il* in French depending on the gender of the nouns the pronoun is referring to. In the example above the gender of the noun *énergie* is feminine so we choose the pronoun *elle*. We included the Stanford Coreference Resolution system (Lee et al., 2013) in our model for this reason. Moreover, in an effort to compare the effectiveness of the word embeddings and handcrafted features for capturing morpho-syntactic information, we decided to use Morfette (Seddah et al., 2010) to supply information on gender and number for each French word explicitly.

2 Linguistic analysis and feature selection

We explained above that pronouns agree in gender and number with their antecedents. But apart from gender and number, there are many other factors at play. For example, there are cases where the English pronoun *they* is translated with *on*. This is usually the case when the antecedents of the pronoun are indefinite or even absent. An example from the training data is *someone can grab your ear and say what* ***they*** *have to say.* It is translated in French as *on peut attraper votre oreille et dire ce que l'* ***on*** *a à dire*.

The same happens when there is a passive in English with the pronoun *it* that is translated in French with active voice. The phrase ***It*** *was called* is translated in French with ***On*** *l' a appelé*.

It can also be translated to *il*. For example, when we find a dummy or expletive pronoun in combination with certain classes of verbs such as *pleuvoir* 'rain', *neiger* 'snow', but also with the verb *sembler* 'seem' and *être* 'be' in expressions such as ***It*** *is time to* translated to ***Il*** *est temps de*.

We could go on explaining the linguistic generalities that were attested in the training data. In summary, we concluded that most of the factors will be captured by including the following features :

1. The English pronoun. This will capture the nature of the English pronoun : is it *it* or *they*.
2. Three words in front of and three words after the English pronoun. This will capture whether the passive is used, whether we find one of the verbs that are often found with expletive pronouns etc.
3. Two words in front of and three words after the French pronoun. This will capture whether we find active or passive voice in French, whether we find one of the verbs that are often found with expletive pronouns and so on.
4. Antecedents and postcedents of the French pronoun. This will capture whether there are antecedents at all and if they are found how definite they are. We can also infer the gender of the antecedents to determine whether to use masculine of feminine forms of pronouns in French.

3 The neural network classifier

3.1 Concept

The neural network structure is described in figure 1. Overall, it resembles the feed-forward neural network structure used in the Continuous Space Translation Model (Son et al., 2012), in which the input layer contains the English words on the source side, and the French words on the target side of the bitext. By using the toolkit learning the word vectors, known as word2vec (Mikolov et al., 2013a) in Python (Řehůřek and Sojka, 2010), we trained two different projection matrices for English and French correspondingly.

A conventional Multi-Layer Perceptron (MLP) on top of the distributed representations maps the input sequences into the pronouns. It is notable that this task is much simpler than the concept used in language models and translation models, in which the output layer needs to be hierachically organised to deal with the gigantic size of the vocabulary. This pronoun prediction task only needs to deal with several pronouns of the target language. If the feature set is limited to only the target words (French), the model is almost identical to a mini language model learning the probabilities of the long n-grams predicting the pronoun class.

In order to include additional features such as the antecedents of the pronoun or morphological tags of the French words, we extend the input lay-

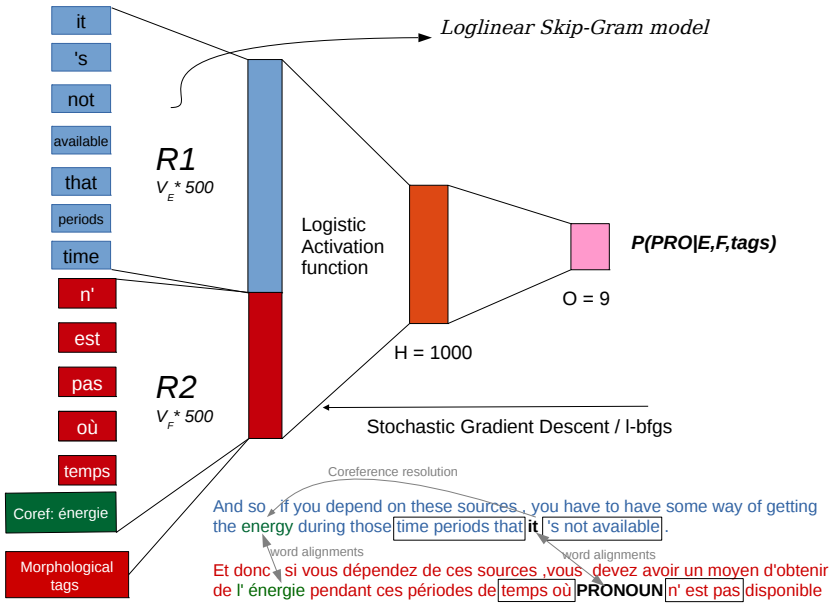

FIGURE 1 – The overview structure of the neural network classifier. The words are transformed into 500-size vectors with two projection matrices $R1$ and $R2$. The size of the hidden layer is 1000, while the output layer gives probabilities for 9 outputs.

ers with additional vectors. One difficulty of the corefence resolution is that the co-referring noun phrases have inconsistent length and might contain a headword and possibly determiners, adjectives or adverbs along with the headword. Our solution was to take only the French words aligned to the English headwords found by the coreference locator [2] for the feature. Therefore, the whole antecedent phrase representation is the average of the French word vectors composing that phrase. That vector is then concatenated to the total feature vector. An additional difficulty lies in the fact that there might be several co-referring expressions for one pronoun, we therefore averaged the projected vectors of all co-referring headwords as done in Hardmeier et al. (2013), but without the probability weighting, since the Stanford Coreference Resolution system does not provide such probabilities.

Technically, this feature is not fully utilised. Coreference resolution can only be found on 30% of the samples. Due to our time and resource limit, we only managed to investigate the antecedents by looking backward one sentence. There are samples whose antecedents are the pronouns of the previous sentences, rendering the feature useless.

As for the linguistic annotations for French morphological features, we treated the tags as one-hot vectors with the size as the total number of tags. Each tagged word is then converted to a corresponding vector, which is then integrated to the input layer of the MLP (the output of the projection layer in the figure). A similar approach was chosen for including the morphological tags of the antecedents as features, where we took the tag vectors of all words in the head-phrases and concatenate the averaged one into the ultimate feature vector.

3.2 Training

As described in the introduction, we trained the system using two separate processes :

- Training the word2vec for continuous representation of English and French words
- Training the MLP classifier

The first part of the training is performed by following the log-linear model concepts proposed by Mikolov et al. (2013a). Fundamentally, word regularities are learnt by using a log-linear classifier to predict a particular word based on its surrounding words (Continuous Bag-Of-Words approach) or to predict the surrounding words based on the current words (Skip-gram approach).

The neural network classifier is trained in order to maximize the log-likelihood of the training data. Backward propagation with Stochastic Gradient Descent optimisation process is performed to

2. These include referential links within the same sentence.

obtain the model weights.

Notably, training the neural network is more demanding than training the word vectors, from a similar amount of data. Consequently, compared to the original training scheme used in language models, we are able to utilise more data for training the word vectors, thus covering a larger vocabulary than the training data provided as the bitext. The difference of the training data for the two parts, as well the parameter selection will be described in the subsequent section.

4 Experiment Setups

4.1 Corpora

The organisers provided us with three different corpora :

- The TED (IWSLT2014) corpus containing approximately 179k bi-sentences.
- The News Commentary corpus, with around 180k sentence pairs.
- The Europarl dataset, originally collected by Koehn (2005) having 2 million sentences.

All three datasets are employed for training the word vectors. Specifically, the projection matrix for each language is trained from approximately 100 million words, comprised of 20k size vocabularies. For training the MLP, we ran experiments with only the in-domain data (TED). For the final submissions of the task, we include another system trained with a larger set of data, including the TED and News Commentary corpora.

4.2 Word2vec training

Regarding architectures, since it is known in previous research (Mikolov et al., 2013a) that the Skip-gram architecture is dominating in terms of modeling the semantics of words, while the CBOW structure is better at capturing morpho-syntactic regularities, we experimented with both architectures to train the projection matrices.

Two important parameters in word2vec are negative sampling and sub-sampling. Negative sampling alters the objective function, from maximizing the corpus probability, that is from the conditional probabilities of the context words given the input words to maximizing directly the quality of the word representations, related to the joint probability of the words and the contexts (Mikolov et al., 2013b; Goldberg and Levy, 2014). "Negative Sample" indicates that, for each sample of word/context, k other samples are drawn randomly assuming they are all negative. The optimisation process only concerns the word representations, rather than the data likelihood. The k value used to generate negative samples is 10 in our setup, which is recommeneded for our corpus size in previous works (Mikolov et al., 2013b).

Sub-sampling is the act of downsampling the very frequent words, based on the intuition that distributional vectors of those words do not change much throughout the training data, plus they do not hold useful information. When sub-sampling was set to 10^{-5}, the performance on the development data was considerably reduced so we decided to leave it out for the remainder of the experiments. As we stated before, it is possible that the frequent words, such as determiners, are necessary for the task.

Our experiments were conducted to observe the impact of word vectors serving pronoun translation, using negative sampling or hierarchical softmax (which is the training method used when negative sampling is disabled). Besides, the context of each word is chosen as 10 (5 words per side). The learned vectors have the size of 500, which are 40 times smaller than the vocabularies.

4.3 Neural Network training

As aforementioned, there are three types of features fed into the MLP :

- Context words from the source side and target side of the translation. Their vectors are treated as the input of the MLP.
- The search for antecedents was performed by the Stanford Coreference Resolution system (Lee et al., 2013). The English words corresponding to the French placeholder are found based on the alignments. Afterwards, coreference resolution is done on the English side by backtracking one sentence and the word alignments help us map the English antecedents to the French counterpart. The feature for the MLP is eventually the averaged word vectors of all words in the French antecedents.
- The Morfette morphological analyser (Seddah et al., 2010) is used to tag each French word with morphological labels, indicating their number and gender properties. We represent such properties as one-hot vectors, showing the index of the tag in the tag list, whose size is 97.

Due to time limitation, we chose to tune the hyper parameters of the network by using the development data. The result of this tuning process is that the activation function is **logistic**, the training algorithm is **l-bfgs** and the hidden layer size is 1000. The experiments were conducted with the Scikit-Learn tool kit (Pedregosa et al., 2011).

5 Results

5.1 Architecture and Feature effect

TABLE 1 – *Results on development set, in macro-average F-measure (%). Comparison of features (English words(E), Coreference(C), French words (F) and Morphological tags(M)), Skip-gram and CBOW architectures, trained with Hierarchical Softmax (HS) and Negative Sampling (NS).*

Features	word2vec Architecture			
	Skip-gram		CBOW	
	HS	NS	HS	NS
English words	37.6	32.6	36.0	32.7
E+Coreference	38.2	32.9	38.0	31.9
E+C+C_MorpTags	39.2	32.7	36.1	34.7
E+C+C_M+French w.	58.4	43.1	58.1	40.6
E+C+C_M+F+F_M	**64.8**	49.7	57.2	50.0

The experimental results for feature engineering and model variations are summerised in Table 1. In total, we exploited 5 progressive feature sets, testing them with two word2vec architectures (Skip-gram and CBOW), each of which is trained with two different methods : Hierarchical Softmax (HS) and Negative Sampling (NS).

Regarding features, the antecedents are shown to be little informative. We see two main reasons for this. First, we explained in Section 3 that we implemented coreference resolution in a suboptimal way due to time restrictions. We will show in the error analysis that the largest part of the mistakes are due to suboptimal coreference handling. Second, in the setup provided for this shared task, the words surrounding the placeholder provide gender and number information already. This fact will downplay the added value of coreference resolution. In an ideal setting the context words would have been normalised. The French words, as expected, contributed greatly for the classification task. They capture many of the linguistic regularities described in Section 2 and on top of that, they often provide gender and number information in the given task setup.

Looking at the difference between the two training methods for both word2vec architectures, the word vectors trained by negative sampling surprisingly fell behind the ones with hierarchical softmax. With the best feature set (E+C+C_M+F+F_M), the HS models outperformed the NS ones by nearly 20% relatively. The reason why NS was effective in previous research is unknown (Goldberg and Levy, 2014), yet it is possible that the dataset in our experiment is preferable for HS in terms of size.

Lastly, we want to discuss the difference in ability of Skip-gram and CBOW models to capture semantic versus morpho-syntactic regularities. From Table 1, we can infer that the CBOW model is able to capture morpho-syntactic regularities, which Skip-gram cannot, which is in line with previous work (Mikolov et al., 2013a). For the Skip-gram models (for the better word vectors trained with HS), the addition of the Morfette tags always led to improvement, especially with the tags of the surrounding French words. The scenario is reversed for the CBOW models, where adding the morphological tags decreased the performance of the system (HS). On the other hand, no matter how well CBOW captures the morpho-syntactic regularities, it falls short in general as Skip-gram outperforms it in all settings (HS). In this task that requires both semantic and morpho-syntactic information, we are best off with a superior semantic model (Skip-gram) in combination with an external tool for morphological analysis.

5.2 Final results on test set

For the final submission on the test set provided by the shared task organisers, we employed the final setting consisting of the best feature set, with the word vectors trained with Skip-gram architecture and hierarchical softmax optimisation, which delivered the highest F-measure for the development set. Furthermore, we doubled the amount of training data, by adding the News Commentary corpus into the training data.

We report results for both fine-grained evaluation (9 classes) and the coarse-grained evaluation (7 classes) as provided by the official scorer. As can be seen in the comparative evaluations provided by the overview paper (Hardmeier et al., 2015), our system is in the top-three in the fine-grained evaluation. A closer look at the performance per class across systems shows that our sys-

tem has particular problems keeping *cela* and *ça* apart, with an F-measure as low as 7.1% for *cela*. We will argue in the error analysis, that we found this distinction to be quite arbitrary in the given data. In the coarse-grained evaluation provided, in which *cela* has been merged with *ça* and *on* has been merged with OTHER, we outperform all competing systems.

TABLE 2 – *Results on test set with additional training data, in macro-average F-measure for both the fine-grained evaluation (9 classes) and the coarse-grained evaluation (7 classes) (%).*

Training data	Fine-g. eval.	Coarse-g. eval.
TED	56.1	**65.8**
TED + NC	**56.5**	65.4

The performance difference between the development set and the test set are large. Although we did not find a clear reason for why this is the case, we point to the overview paper that shows that the baseline also performs very differently on the two sets. They attribute this effect to the test set's better coverage of infrequent pronouns.

Adding more training data does not lead to clear improvements. One reason for that seems to be that the class distribution of the out-of-domain data is rather different from the in-domain data.

5.3 Error analysis

We inspected the output of our best system on the development data in order to find the major sources of error. We randomly selected about 2/3rd of the data. We came to the following conclusions : The model manages to capture the linguistic regularities described in Section 2 rather well. It does less well on capturing the antecedent and using this type of information for predicting the French pronoun. Approximately 50% of the errors made by our system seemed due to an improper handling of coreference. We explained that our implementation of features for coreference was suboptimal, but improving this component to handle coreference perfectly is very hard as shown in previous work (Hardmeier et al., 2013). The coreference needs to be transferred from the English to the French sentences and alignment errors are added to mistakes already present in the original English coreference chains.
On the bright side of things, we saw that approximately 10% of the errors were in fact perfectly acceptable. For example, the difference between *ça* and *cela* is merely due to differences is register, and we saw individual speakers switching back and forth between the two in one conversation. The coarse-grained evaluation proposed conflates *ça* and *cela*.

6 Conclusions

In this paper, we described a system that addresses the task of cross-lingual pronoun prediction from English to French. We show that it is a hard task that requires many different types of information, such as morpho-syntactic information as well as semantics of context words and identification of antecedents of the French pronoun.

We proposed a model that captures linguistic generalisation using word embeddings that are fed into a MLP in addition to morphological analysis and coreference resolution. Although word embeddings (CBOW) are known to capture morpho-syntactic operations quite well, we show that using a standalone morphological analyser in combination with the semantically stronger version of the continuous word space models (Skip-gram) produces the best results (56.5% on the test set). Coreference resolution showed the least beneficial in our experiments. This seems due to the suboptimal implementation of this type of information in our model and the gender and number information contained in the French context words.

The error analysis showed that half of the errors could be solved with a proper implementation of coreference resolution, which is however not trivial to do. 10% percent of the errors were in fact acceptable variations. The coarse-grained evaluation proposed conflates some of these seemingly equivalent classes and results in a 65.4%, the best score reported by participating teams. Also, performance numbers should be higher, when based on human judgements.

Acknowledgments

This research was funded by the Erasmus Mundus European Masters Program in Language and Communication Technologies (LCT) and the University of Malta and has been carried out using computational facilities procured through the European Regional Development Fund, Project ERDF-080 'A Supercomputing Laboratory for the University of Malta' (`http://www.um.edu.mt/research/scienceeng/erdf_080`).

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3 :1137–1155.

Yoav Goldberg and Omer Levy. 2014. word2vec explained : deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv :1402.3722*.

Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013. Latent anaphora resolution for cross-lingual pronoun prediction. In *EMNLP 2013 ; Conference on Empirical Methods in Natural Language Processing ; 18-21 October 2013 ; Seattle, WA, USA*, pages 380–391. Association for Computational Linguistics.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction : Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, Lisbon, Portugal.

Philipp Koehn. 2005. Europarl : A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. 2015. Multilingual reliability and "semantic" structure of continuous word spaces. *IWCS 2015*, page 40.

Hai-Son Le, Ilya Oparin, Alexandre Allauzen, Jean-Luc Gauvain, and François Yvon. 2011. Structured output layer neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5524–5527. IEEE.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4) :885–916.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn : Machine learning in python. *The Journal of Machine Learning Research*, 12 :2825–2830.

Holger Schwenk. 2007. Continuous space language models. *Computer Speech & Language*, 21(3) :492–518.

Djamé Seddah, Grzegorz Chrupała, Özlem Çetinoğlu, Josef Van Genabith, and Marie Candito. 2010. Lemmatization and lexicalized statistical parsing of morphologically rich languages : the case of french. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 85–93. Association for Computational Linguistics.

Le Hai Son, Alexandre Allauzen, and François Yvon. 2012. Continuous space translation models with neural networks. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics : Human language technologies*, pages 39–48. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. *Proceedings of the International Conference on Language Resources and Evaluation*.