Latent Anaphora Resolution for Cross-Lingual Pronoun Prediction

Christian Hardmeier Jörg Tiedemann Joakim Nivre
Uppsala University
Department of Linguistics and Philology
Box 635, 751 26 Uppsala, Sweden
firstname.lastname@lingfil.uu.se

Abstract

This paper addresses the task of predicting the correct French translations of third-person subject pronouns in English discourse, a problem that is relevant as a prerequisite for machine translation and that requires anaphora resolution. We present an approach based on neural networks that models anaphoric links as latent variables and show that its performance is competitive with that of a system with separate anaphora resolution while not requiring any coreference-annotated training data. This demonstrates that the information contained in parallel bitexts can successfully be used to acquire knowledge about pronominal anaphora in an unsupervised way.

1 Motivation

When texts are translated from one language into another, the translation reconstructs the meaning or function of the source text with the means of the target language. Generally, this has the effect that the entities occurring in the translation and their mutual relations will display similar patterns as the entities in the source text. In particular, coreference patterns tend to be very similar in translations of a text, and this fact has been exploited with good results to project coreference annotations from one language into another by using word alignments (Postolache et al., 2006; Rahman and Ng, 2012).

On the other hand, what is true in general need not be true for all types of linguistic elements. For instance, a substantial percentage of the English third-person subject pronouns *he*, *she*, *it* and *they* does not get realised as pronouns in French translations (Hardmeier, 2012). Moreover, it has been recognised by various authors in the statistical machine translation (SMT) community (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012) that pronoun translation is a difficult problem because, even when a pronoun does get translated as a pronoun, it may require choosing the correct word form based on agreement features that are not easily predictable from the source text.

The work presented in this paper investigates the problem of cross-lingual pronoun prediction for English-French. Given an English pronoun and its discourse context as well as a French translation of the same discourse and word alignments between the two languages, we attempt to predict the French word aligned to the English pronoun. As far as we know, this task has not been addressed in the literature before. In our opinion, it is interesting for several reasons. By studying pronoun prediction as a task in its own right, we hope to contribute towards a better understanding of pronoun translation with a long-term view to improving the performance of SMT systems. Moreover, we believe that this task can lead to interesting insights about anaphora resolution in a multi-lingual context. In particular, we show in this paper that the pronoun prediction task makes it possible to model the resolution of pronominal anaphora as a latent variable and opens up a way to solve a task relying on anaphora resolution without using any data annotated for anaphora. This is what we consider the main contribution of our present work.

We start by modelling cross-lingual pronoun prediction as an independent machine learning task after doing anaphora resolution in the source language (English) using the BART software (Broscheit et al., 2010). We show that it is difficult to achieve satisfactory performance with standard maximum-

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 380–391,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

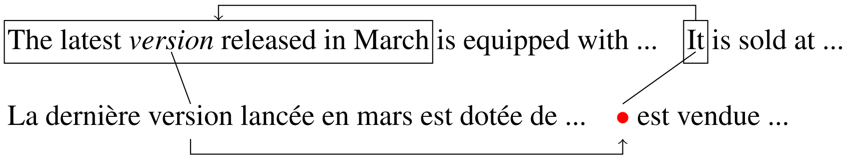
The latest *version* released in March is equipped with ... It is sold at ...
La dernière version lancée en mars est dotée de ... ● est vendue ...

Figure 1: Task setup

entropy classifiers especially for low-frequency pronouns such as the French feminine plural pronoun *elles*. We propose a neural network classifier that achieves better precision and recall and manages to make reasonable predictions for all pronoun categories in many cases.

We then go on to extend our neural network architecture to include anaphoric links as latent variables. We demonstrate that our classifier, now with its own source language anaphora resolver, can be trained successfully with backpropagation. In this setup, we no longer use the machine learning component included in the external coreference resolution system (BART) to predict anaphoric links. Anaphora resolution is done by our neural network classifier and requires only some quantity of word-aligned parallel data for training, completely obviating the need for a coreference-annotated training set.

2 Task Setup

The overall setup of the classification task we address in this paper is shown in Figure 1. We are given an English discourse containing a pronoun along with its French translation and word alignments between the two languages, which in our case were computed automatically using a standard SMT pipeline with GIZA++ (Och and Ney, 2003). We focus on the four English third-person subject pronouns *he*, *she*, *it* and *they*. The output of the classifier is a multinomial distribution over six classes: the four French subject pronouns *il*, *elle*, *ils* and *elles*, corresponding to masculine and feminine singular and plural, respectively; the impersonal pronoun *ce/c'*, which occurs in some very frequent constructions such as *c'est* (*it is*); and a sixth class OTHER, which indicates that none of these pronouns was used. In general, a pronoun may be aligned to multiple words; in this case, a training example is counted as a positive example for a class if the target word occurs among the words aligned to the pronoun, irrespective of the presence of other aligned tokens.

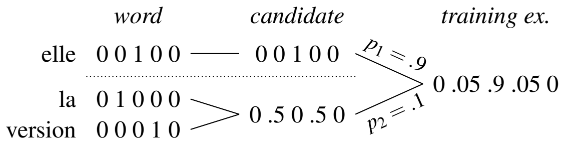

Figure 2: Antecedent feature aggregation

This task setup resembles the problem that an SMT system would have to solve to make informed choices when translating pronouns, an aspect of translation neglected by most existing SMT systems. An important difference between the SMT setup and our own classifiers is that we use context from human-made translations for prediction. This potentially makes the task both easier and more difficult; easier, because the context can be relied on to be correctly translated, and more difficult, because human translators frequently create less literal translations than an SMT system would. Integrating pronoun prediction into the translation process would require significant changes to the standard SMT decoding setup in order to take long-range dependencies in the target language into account, which is why we do not address this issue in our current work.

In all the experiments presented in this paper, we used features from two different sources:

- *Anaphora context features* describe the source language pronoun and its immediate context consisting of three words to its left and three words to its right. They are encoded as vectors whose dimensionality is equal to the source vocabulary size with a single non-zero component indicating the word referred to (one-hot vectors).
- *Antecedent features* describe an antecedent candidate. Antecedent candidates are represented by the target language words aligned to the syntactic head of the source language markable

	TED	News
ce	16.3 %	6.4 %
elle	7.1 %	10.1 %
elles	3.0 %	3.9 %
il	17.1 %	26.5 %
ils	15.6 %	15.1 %
OTHER	40.9 %	38.0 %

Table 1: Distribution of classes in the training data

noun phrase as identified by the Collins head finder (Collins, 1999).

The different handling of anaphora context features and antecedent features is due to the fact that we always consider a constant number of context words on the source side, whereas the number of word vectors to be considered depends on the number of antecedent candidates and on the number of target words aligned to each antecedent.

The encoding of the antecedent features is illustrated in Figure 2 for a training example with two antecedent candidates translated to *elle* and *la version*, respectively. The target words are represented as one-hot vectors with the dimensionality of the target language vocabulary. These vectors are then averaged to yield a single vector per antecedent candidate. Finally, the vectors of all candidates for a given training example are weighted by the probabilities assigned to them by the anaphora resolver (p_1 and p_2) and summed to yield a single vector per training example.

3 Data Sets and External Tools

We run experiments with two different test sets. The *TED* data set consists of around 2.6 million tokens of lecture subtitles released in the WIT^3 corpus (Cettolo et al., 2012). The WIT^3 training data yields 71,052 examples, which were randomly partitioned into a training set of 63,228 examples and a test set of 7,824 examples. The official WIT^3 development and test sets were not used in our experiments. The *news-commentary* data set is version 6 of the parallel news-commentary corpus released as a part of the WMT 2011 training data[1]. It contains around 2.8 million tokens of news text and yields 31,017 data points, which were randomly split into 27,900 training examples and 3,117 test instances. The distribution of the classes in the two training sets is shown in Table 1. One thing to note is the dominance of the OTHER class, which pools together such different phenomena as translations with other pronouns not in our list (e. g., *celui-ci*) and translations with full noun phrases instead of pronouns. Splitting this group into more meaningful subcategories is not straightforward and must be left to future work.

The feature setup of all our classifiers requires the detection of potential antecedents and the extraction of features pairing anaphoric pronouns with antecedent candidates. Some of our experiments also rely on an external anaphora resolution component. We use the open-source anaphora resolver BART to generate this information. BART (Broscheit et al., 2010) is an anaphora resolution toolkit consisting of a markable detection and feature extraction pipeline based on a variety of standard natural language processing (NLP) tools and a machine learning component to predict coreference links including both pronominal anaphora and noun-noun coreference. In our experiments, we always use BART's markable detection and feature extraction machinery. Markable detection is based on the identification of noun phrases in constituency parses generated with the Stanford parser (Klein and Manning, 2003). The set of features extracted by BART is an extension of the widely used mention-pair anaphora resolution feature set by Soon et al. (2001) (see below, Section 6).

In the experiments of the next two sections, we also use BART to predict anaphoric links for pronouns. The model used with BART is a maximum entropy ranker trained on the *ACE02-npaper* corpus (LDC2003T11). In order to obtain a probability distribution over antecedent candidates rather than one-best predictions or coreference sets, we modified the ranking component with which BART resolves pronouns to normalise and output the scores assigned by the ranker to all candidates instead of picking the highest-scoring candidate.

4 Baseline Classifiers

In order to create a simple, but reasonable baseline for our task, we trained a maximum entropy (ME)

[1] `http://www.statmt.org/wmt11/translation-task.html` (3 July 2013).

	TED		
	(Accuracy: 0.685)		
	P	R	F
ce	0.593	0.728	0.654
elle	0.798	0.523	0.632
elles	0.812	0.164	0.273
il	0.764	0.550	0.639
ils	0.632	0.949	0.759
OTHER	0.724	0.692	0.708

	News commentary		
	(Accuracy: 0.576)		
	P	R	F
ce	0.508	0.294	0.373
elle	0.530	0.312	0.393
elles	0.538	0.062	0.111
il	0.600	0.666	0.631
ils	0.593	0.769	0.670
OTHER	0.564	0.609	0.586

Table 2: Maximum entropy classifier results

	TED		
	(Accuracy: 0.700)		
	P	R	F
ce	0.634	0.747	0.686
elle	0.756	0.617	0.679
elles	0.679	0.319	0.434
il	0.719	0.591	0.649
ils	0.663	0.940	0.778
OTHER	0.743	0.678	0.709

	News commentary		
	(Accuracy: 0.576)		
	P	R	F
ce	0.477	0.344	0.400
elle	0.498	0.401	0.444
elles	0.565	0.116	0.193
il	0.655	0.626	0.640
ils	0.570	0.834	0.677
OTHER	0.567	0.573	0.570

Table 3: Neural network classifier with anaphoras resolved by BART

classifier with the MegaM software package[2] using the features described in the previous section and the anaphora links found by BART. Results are shown in Table 2. The baseline results show an overall higher accuracy for the TED data than for the news-commentary data. While the precision is above 50 % in all categories and considerably higher in some, recall varies widely.

The pronoun *elles* is particularly interesting. This is the feminine plural of the personal pronoun, and it usually corresponds to the English pronoun *they*, which is not marked for gender. In French, *elles* is a marked choice which is only used if the antecedent exclusively refers to females or feminine-gendered objects. The presence of a single item with masculine grammatical gender in the antecedent will trigger the use of the masculine plural pronoun *ils* instead. This distinction cannot be predicted from the English source pronoun or its context; making correct predictions requires knowledge about the antecedent of the pronoun. Moreover, *elles* is a low-frequency pronoun. There are only 1,909 occurrences of this pronoun in the TED training data, and 1,077 in the news-commentary training set. Because of these special properties of the feminine plural class, we argue that the performance of a classifier on *elles* is a good indicator of how well it can represent relevant knowledge about pronominal anaphora as opposed to overfitting to source contexts or acting on prior assumptions about class frequencies.

In accordance with the general linguistic preference for *ils*, the classifier tends to predict *ils* much more often than *elles* when encountering an English plural pronoun. This is reflected in the fact that *elles* has much lower recall than *ils*. Clearly, the classifier achieves a good part of its accuracy by making majority choices without exploiting deeper knowledge about the antecedents of pronouns.

An additional experiment with a subset of 27,900 training examples from the TED data confirms that the difference between TED and news commentaries is not just an effect of training data size, but that TED data is genuinely easier to predict than news commentaries. In the reduced data TED condition, the classifier achieves an accuracy of 0.673. Precision and recall of all classifiers are much closer to the

[2] `http://www.umiacs.umd.edu/~hal/megam/` (20 June 2013).

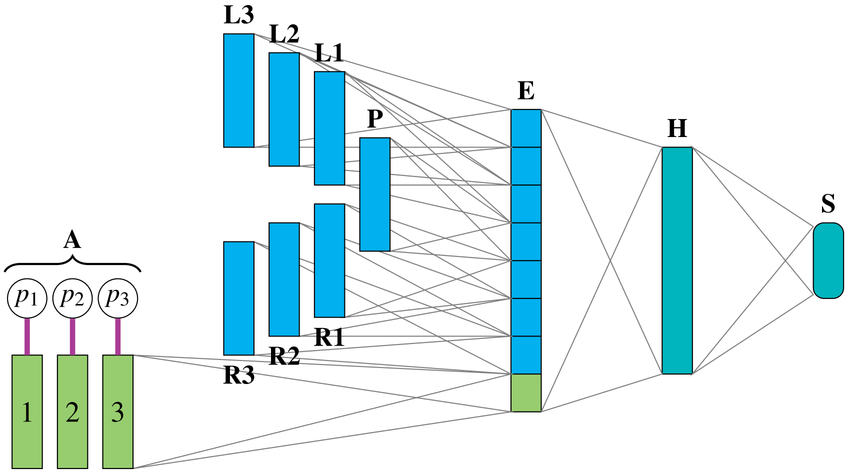

Figure 3: Neural network for pronoun prediction

large-data TED condition than to the news commentary experiments, except for *elles*, where we obtain an F-score of 0.072 (P 0.818, R 0.038), indicating that small training data size is a serious problem for this low-frequency class.

5 Neural Network Classifier

In the previous section, we saw that a simple multi-class maximum entropy classifier, while making correct predictions for much of the data set, has a significant bias towards making majority class decisions, relying more on prior assumptions about the frequency distribution of the classes than on antecedent features when handling examples of less frequent classes. In order to create a system that can be trained to rely more explicitly on antecedent information, we created a neural network classifier for our task. The introduction of a hidden layer should enable the classifier to learn abstract concepts such as gender and number that are useful across multiple output categories, so that the performance of sparsely represented classes can benefit from the training examples of the more frequent classes.

The overall structure of the network is shown in Figure 3. As inputs, the network takes the same features that were available to the baseline ME classifier, based on the source pronoun (**P**) with three words of context to its left (**L1** to **L3**) and three words to its right (**R1** to **R3**) as well as the words aligned to the syntactic head words of all possible antecedent candidates as found by BART (**A**). All words are encoded as one-hot vectors whose dimensionality is equal to the vocabulary size. If multiple words are aligned to the syntactic head of an antecedent candidate, their word vectors are averaged with uniform weights. The resulting vectors for each antecedent are then averaged with weights defined by the posterior distribution of the anaphora resolver in BART (p_1 to p_3).

The network has two hidden layers. The first layer (**E**) maps the input word vectors to a low-dimensional representation. In this layer, the embedding weights for all the source language vectors (the pronoun and its 6 context words) are tied, so if two words are the same, they are mapped to the same lower-dimensional embedding irrespective of their position relative to the pronoun. The embedding of the antecedent word vectors is independent, as these word vectors represent target language words. The entire embedding layer is then mapped to another hidden layer (**H**), which is in turn connected to a softmax output layer (**S**) with 6 outputs representing the classes *ce*, *elle*, *elles*, *il*, *ils* and OTHER. The non-linearity of both hidden layers is the logistic sigmoid function, $f(x) = 1/(1+e^{-x})$.

In all experiments reported in this paper, the dimensionality of the source and target language word embeddings is 20, resulting in a total embedding layer size of 160, and the size of the last hidden layer is equal to 50. These sizes are fairly small. In experiments with larger layer sizes, we were able to obtain similar, but no better results.

The neural network is trained with mini-batch stochastic gradient descent with backpropagated gradients using the RMSPROP algorithm with cross-entropy as the objective function.[3] In contrast to standard gradient descent, RMSPROP normalises the magnitude of the gradient components by dividing them by a root-mean-square moving average. We found this led to faster convergence. Other features of our training algorithm include the use of momentum to even out gradient oscillations, adaptive learning rates for each weight as well as adaptation of the global learning rate as a function of current training progress. The network is regularised with an ℓ_2 weight penalty. Good settings of the initial learning rate and the weight cost parameter (both around 0.001 in most experiments) were found by manual experimentation. Generally, we train our networks for 300 epochs, compute the validation error on a held-out set of some 10 % of the training data after each epoch and use the model that achieved the lowest validation error for testing.

Since the source context features are very informative and it is comparatively more difficult to learn from the antecedents, the network sometimes had a tendency to overfit to the source features and disregard antecedent information. We found that this problem can be solved effectively by presenting a part of the training without any source features, forcing the network to learn from the information contained in the antecedents. In all experiments in this paper, we zero out all source features (input layers **P**, **L1** to **L3** and **R1** to **R3**) with a probability of 50 % in each training example. At test time, no information is zeroed out.

Classification results with this network are shown in Table 3. We note that the accuracy has increased slightly for the TED test set and remains exactly the same for the news commentary corpus. However, a closer look on the results for individual classes reveals that the neural network makes better predictions for almost all classes. In terms of F-score, the only class that becomes slightly worse is the OTHER class for the news commentary corpus because of lower recall, indicating that the neural network classifier is less biased towards using the uninformative OTHER category. Recall for *elle* and *elles* increases considerably, but especially for *elles* it is still quite low. The increase in recall comes with some loss in precision, but the net effect on F-score is clearly positive.

6 Latent Anaphora Resolution

Considering Figure 1 again, we note that the bilingual setting of our classification task adds some information not available to the monolingual anaphora resolver that can be helpful when determining the correct antecedent for a given pronoun. Knowing the gender of the translation of a pronoun limits the set of possible antecedents to those whose translation is morphologically compatible with the target language pronoun. We can exploit this fact to learn how to resolve anaphoric pronouns without requiring data with manually annotated anaphoric links.

To achieve this, we extend our neural network with a component to predict the probability of each antecedent candidate to be the correct antecedent (Figure 4). The extended network is identical to the previous version except for the upper left part dealing with anaphoric link features. The only difference between the two networks is the fact that anaphora resolution is now performed by a part of our neural network itself instead of being done by an external module and provided to the classifier as an input.

In this setup, we still use some parts of the BART toolkit to extract markables and compute features. However, we do not make use of the machine learning component in BART that makes the actual predictions. Since this is the only component trained on coreference-annotated data in a typical BART configuration, no coreference annotations are used anywhere in our system even though we continue to rely on the external anaphora resolver for preprocessing to avoid implementing our own markable and feature extractors and to make comparison easier.

For each candidate markable identified by BART's preprocessing pipeline, the anaphora resolution model receives as input a link feature vector (**T**) describing relevant aspects of the antecedent candidate-anaphora pair. This feature vector is generated by the feature extraction machinery in BART and includes a standard feature set for coreference resolution partially based on work by Soon et al. (2001). We use the following feature extractors in BART, each of

[3] Our training procedure is greatly inspired by a series of online lectures held by Geoffrey Hinton in 2012 (`https://www.coursera.org/course/neuralnets`, 10 September 2013).

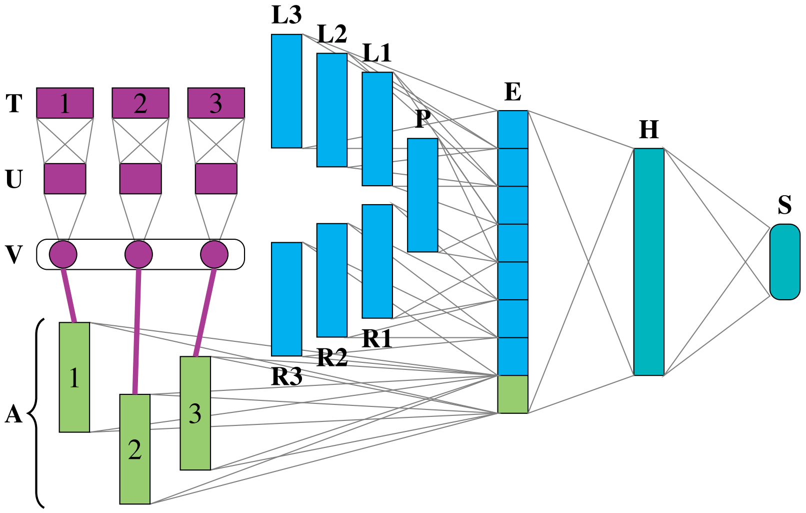

Figure 4: Neural network with latent anaphora resolution

which can generate multiple features:

- Anaphora mention type
- Gender match
- Number match
- String match
- Alias feature (Soon et al., 2001)
- Appositive position feature (Soon et al., 2001)
- Semantic class (Soon et al., 2001)
- Semantic class match
- Binary distance feature
- Antecedent is first mention in sentence

Our baseline set of features was borrowed wholesale from a working coreference system and includes some features that are not relevant to the task at hand, e. g., features indicating that the anaphora is a pronoun, is not a named entity, etc. After removing all features that assume constant values in the training set when resolving antecedents for the set of pronouns we consider, we are left with a basic set of 37 anaphoric link features that are fed as inputs to our network. These features are exactly the same as those available to the anaphora resolution classifier in the BART system used in the previous section.

Each training example for our network can have an arbitrary number of antecedent candidates, each of which is described by an antecedent word vector (**A**) and by an anaphoric link vector (**T**). The anaphoric link features are first mapped to a regular hidden layer with logistic sigmoid units (**U**). The activations of the hidden units are then mapped to a single value, which functions as an element in a softmax layer over all antecedent candidates (**V**). This softmax layer assigns a probability to each antecedent candidate, which we then use to compute a weighted average over the antecedent word vector, replacing the probabilities p_i in Figures 2 and 3.

At training time, the network's anaphora resolution component is trained in exactly the same way as the rest of the network. The error signal from the embedding layer is backpropagated both to the weight matrix defining the antecedent word embedding and to the anaphora resolution subnetwork. Note that the number of weights in the network is the same for all training examples even though the number of antecedent candidates varies because all weights related to antecedent word features and anaphoric link features are shared between all antecedent candidates.

One slightly uncommon feature of our neural network is that it contains an internal softmax layer to generate normalised probabilities over all possible antecedent candidates. Moreover, weights are shared between all antecedent candidates, so the inputs of our internal softmax layer share dependencies on the same weight variables. When computing derivatives with backpropagation, these shared dependencies must be taken into account. In particular, the outputs y_i of the antecedent resolution layer are the result of a softmax applied to functions of some shared variables q:

$$y_i = \frac{\exp f_i(q)}{\sum_k \exp f_k(q)} \tag{1}$$

The derivatives of any y_i with respect to q, which can be any of the weights in the anaphora resolution subnetwork, have dependencies on the derivatives of the other softmax inputs with respect to q:

$$\frac{\partial y_i}{\partial q} = y_i \left(\frac{\partial f_i(q)}{\partial q} - \sum_k y_k \frac{\partial f_k(q)}{\partial q} \right) \qquad (2)$$

This makes the implementation of backpropagation for this part of the network somewhat more complicated, but in the case of our networks, it has no major impact on training time.

Experimental results for this network are shown in Table 4. Compared with Table 3, we note that the overall accuracy is only very slightly lower for TED, and for the news commentaries it is actually better. When it comes to F-scores, the performance for *elles* improves by a small amount, while the effect on the other classes is a bit more mixed. Even where it gets worse, the differences are not dramatic considering that we eliminated a very knowledge-rich resource from the training process. This demonstrates that it is possible, in our classification task, to obtain good results without using any data manually annotated for anaphora and to rely entirely on unsupervised latent anaphora resolution.

7 Further Improvements

The results presented in the preceding section represent a clear improvement over the ME classifiers in Table 2, even though the overall accuracy increased only slightly. Not only does our neural network classifier achieve better results on the classification task at hand without requiring an anaphora resolution classifier trained on manually annotated data, but it performs clearly better for the feminine categories that reflect minority choices requiring knowledge about the antecedents. Nevertheless, the performance is still not entirely satisfactory.

By subjecting the output of our classifier on a development set to a manual error analysis, we found that a fairly large number of errors belong to two error types: On the one hand, the preprocessing pipeline used to identify antecedent candidates does not always include the correct antecedent in the set presented to the neural network. Whenever this occurs, it is obvious that the classifier cannot possibly find the correct antecedent. Out of 76 examples of the category *elles* that had been mistakenly predicted as *ils*, we found that 43 suffered from this problem. In other classes, the problem seems to be somewhat less common, but it still exists. On the other hand, in many cases (23 out of 76 for the category mentioned before) the anaphora resolution subnetwork does identify an antecedent manually recognised to belong to the right gender/number group, but still predicts an incorrect pronoun. This may indicate that the network has difficulties learning a correct gender/number representation for all words in the vocabulary.

7.1 Relaxing Markable Extraction

The pipeline we use to extract potential antecedent candidates is borrowed from the BART anaphora resolution toolkit. BART uses a syntactic parser to identify noun phrases as markables. When extracting antecedent candidates for coreference prediction, it starts by considering a window consisting of the sentence in which the anaphoric pronoun is located and the two immediately preceding sentences. Markables in this window are checked for morphological compatibility in terms of gender and number with the anaphoric pronoun, and only compatible markables are extracted as antecedent candidates. If no compatible markables are found in the initial window, the window is successively enlarged one sentence at a time until at least one suitable markable is found.

Our error analysis shows that this procedure misses some relevant markables both because the initial two-sentence extraction window is too small and because the morphological compatibility check incorrectly filters away some markables that should have been considered as candidates. By contrast, the extraction procedure does extract quite a number of first and second person noun phrases (*I*, *we*, *you* and their oblique forms) in the TED talks which are extremely unlikely to be the antecedent of a later occurrence of *he*, *she*, *it* or *they*. As a first step, we therefore adjust the extraction criteria to our task by increasing the initial extraction window to five sentences, excluding first and second person markables and removing the morphological compatibility requirement. The compatibility check is still used to control expansion of the extraction window, but it is no longer applied to filter the extracted markables. This increases the accuracy to 0.701 for TED and 0.602 for the news

	TED (Accuracy: 0.696)				News commentary (Accuracy: 0.597)		
	P	R	F		P	R	F
ce	0.618	0.722	0.666	*ce*	0.419	0.368	0.392
elle	0.754	0.548	0.635	*elle*	0.547	0.460	0.500
elles	0.737	0.340	0.465	*elles*	0.539	0.135	0.215
il	0.718	0.629	0.670	*il*	0.623	0.719	0.667
ils	0.652	0.916	0.761	*ils*	0.596	0.783	0.677
OTHER	0.741	0.682	0.711	OTHER	0.614	0.544	0.577

Table 4: Neural network classifier with latent anaphora resolution

	TED (Accuracy: 0.713)				News commentary (Accuracy: 0.626)		
	P	R	F		P	R	F
ce	0.611	0.723	0.662	*ce*	0.492	0.324	0.391
elle	0.749	0.596	0.664	*elle*	0.526	0.439	0.478
elles	0.602	0.616	0.609	*elles*	0.547	0.558	0.552
il	0.733	0.638	0.682	*il*	0.599	0.757	0.669
ils	0.710	0.884	0.788	*ils*	0.671	0.878	0.761
OTHER	0.760	0.704	0.731	OTHER	0.681	0.526	0.594

Table 5: Final classifier results

commentaries, while the performance for *elles* improves to F-scores of 0.531 (TED; P 0.690, R 0.432) and 0.304 (News commentaries; P 0.444, R 0.231), respectively. Note that these and all the following results are not directly comparable to the ME baseline results in Table 2, since they include modifications and improvements to the training data extraction procedure that might possibly lead to benefits in the ME setting as well.

7.2 Adding Lexicon Knowledge

In order to make it easier for the classifier to identify the gender and number properties of infrequent words, we extend the word vectors with features indicating possible morphological features for each word. In early experiments with ME classifiers, we found that our attempts to do proper gender and number tagging in French text did not improve classification performance noticeably, presumably because the annotation was too noisy. In more recent experiments, we just add features indicating all possible morphological interpretations of each word, rather than trying to disambiguate them. To do this, we look up the morphological annotations of the French words in the Lefff dictionary (Sagot et al., 2006) and introduce a set of new binary features to indicate whether a particular reading of a word occurs in that dictionary. These features are then added to the one-hot representation of the antecedent words. Doing so improves the classifier accuracy to 0.711 (TED) and 0.604 (News commentaries), while the F-scores for *elles* reach 0.589 (TED; P 0.649, R 0.539) and 0.500 (News commentaries; P 0.545, R 0.462), respectively.

7.3 More Anaphoric Link Features

Even though the modified antecedent candidate extraction with its larger context window and without the morphological filter results in better performance on both test sets, additional error analysis reveals that the classifiers has greater problems identifying the correct markable in this setting. One reason for this may be that the baseline anaphoric link feature set described above (Section 6) only includes two very rough binary distance features which indicate whether or not the anaphora and the antecedent candidate occur in the same or in immediately adjacent sentences. With the larger context window, this may be too unspecific. In our final experiment, we therefore enable some additional features which are available in BART, but disabled in the baseline system:

- Distance in number of markables
- Distance in number of sentences
- Sentence distance, log-transformed
- Distance in number of words
- Part of speech of head word

Most of these encode the distance between the anaphora and the antecedent candidate in more precise ways. Complete results for this final system are presented in Table 5.

Including these additional features leads to another slight increase in accuracy for both corpora, with similar or increased classifier F-scores for most classes except *elle* in the news commentary condition. In particular, we should like to point out the performance of our benchmark classifier for *elles*, which suffered from extremely low recall in the first classifiers and approaches the performance of the other classes, with nearly balanced precision and recall, in this final system. Since *elles* is a low-frequency class and cannot be reliably predicted using source context alone, we interpret this as evidence that our final neural network classifier has incorporated some relevant knowledge about pronominal anaphora that the baseline ME classifier and earlier versions of our network have no access to. This is particularly remarkable because no data manually annotated for coreference was used for training.

8 Related work

Even though it was recognised years ago that the information contained in parallel corpora may provide valuable information for the improvement of anaphora resolution systems, there have not been many attempts to cash in on this insight. Mitkov and Barbu (2003) exploit parallel data in English and French to improve pronominal anaphora resolution by combining anaphora resolvers for the individual languages with handwritten rules to resolve conflicts between the output of the language-specific resolvers. Veselovská et al. (2012) apply a similar strategy to English-Czech data to resolve different uses of the pronoun *it*. Other work has used word alignments to project coreference annotations from one language to another with a view to training anaphora resolvers in the target language (Postolache et al., 2006; de Souza and Orăsan, 2011). Rahman and Ng (2012) instead use machine translation to translate their test data into a language for which they have an anaphora resolver and then project the annotations back to the original language. Completely unsupervised monolingual anaphora resolution has been approached using, e. g., Markov logic (Poon and Domingos, 2008) and the Expectation-Maximisation algorithm (Cherry and Bergsma, 2005; Charniak and Elsner, 2009). To the best of our knowledge, the direct application of machine learning techniques to parallel data in a task related to anaphora resolution is novel in our work.

Neural networks and deep learning techniques have recently gained some popularity in natural language processing. They have been applied to tasks such as language modelling (Bengio et al., 2003; Schwenk, 2007), translation modelling in statistical machine translation (Le et al., 2012), but also part-of-speech tagging, chunking, named entity recognition and semantic role labelling (Collobert et al., 2011). In tasks related to anaphora resolution, standard feed-forward neural networks have been tested as a classifier in an anaphora resolution system (Stuckardt, 2007), but the network design presented in our work is novel.

9 Conclusion

In this paper, we have introduced cross-lingual pronoun prediction as an independent natural language processing task. Even though it is not an end-to-end task, pronoun prediction is interesting for several reasons. It is related to the problem of pronoun translation in SMT, a currently unsolved problem that has been addressed in a number of recent research publications (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012) without reaching a major breakthrough. In this work, we have shown that pronoun prediction can be effectively modelled in a neural network architecture with relatively simple features. More importantly, we have demonstrated that the task can be exploited to train a classifier with a latent representation of anaphoric links. With parallel text as its only supervision this classifier achieves a level of performance that is similar to, if not better than, that of a classifier using a regular anaphora resolution system trained with manually annotated data.

References

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

Samuel Broscheit, Massimo Poesio, Simone Paolo Ponzetto, Kepa Joseba Rodriguez, Lorenza Romano, Olga Uryupina, Yannick Versley, and Roberto Zanoli. 2010. BART: A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010),* Uppsala, Sweden, 15–16 July 2010.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 148–156, Athens, Greece.

Colin Cherry and Shane Bergsma. 2005. An Expectation Maximization approach to pronoun resolution. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 88–95, Ann Arbor, Michigan.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2461–2505.

José de Souza and Constantin Orăsan. 2011. Can projected chains in parallel corpora help coreference resolution? In Iris Hendrickx, Sobha Lalitha Devi, António Branco, and Ruslan Mitkov, editors, *Anaphora Processing and Applications*, volume 7099 of *Lecture Notes in Computer Science*, pages 59–69. Springer, Berlin.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, France.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289, Paris, France.

Christian Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. *Discours*, 11.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan.

Hai-Son Le, 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, Montréal, Canada.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden.

Ruslan Mitkov and Catalina Barbu. 2003. Using bilingual corpora to improve pronoun resolution. *Languages in Contrast*, 4(2):201–211.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29:19–51.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 650–659, Honolulu, Hawaii.

Oana Postolache, Dan Cristea, and Constantin Orăsan. 2006. Transferring coreference chains through word alignment. In *Proceedings of the 5th Conference on International Language Resources and Evaluation (LREC-2006)*, pages 889–892, Genoa.

Altaf Rahman and Vincent Ng. 2012. Translation-based projection for multilingual coreference resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 720–730, Montréal, Canada.

Benoît Sagot, Lionel Clément, Éric Villemonte de La Clergerie, and Pierre Boullier. 2006. The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of the 5th Conference on International Language Resources and Evaluation (LREC-2006)*, pages 1348–1351, Genoa.

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

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Roland Stuckardt. 2007. Applying backpropagation networks to anaphor resolution. In António Branco, editor, *Anaphora: Analysis, Algorithms and Applications. 6th*

Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007, number 4410 in Lecture Notes in Artificial Intelligence, pages 107–124, Berlin.

Kateřina Veselovská, Ngụy Giang Linh, and Michal Novák. 2012. Using Czech-English parallel corpora in automatic identification of *it*. In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora*, pages 112–120, Istanbul, Turkey.