Hybrid Adaptation of Named Entity Recognition for Statistical Machine Translation

Vassilina NIKOULINA   Agnes SANDOR   Marc DYMETMAN
Xerox Research Center Europe, 6, chemin de Maupertuis,Meylan,FRANCE
`vassilina.nikoulina@xrce.xerox.com, agnes.sandor@xrce.xerox.com,`
`marc.dymetman@xrce.xerox.com`

ABSTRACT

Appropriate Named Entity handling is important for Statistical Machine Translation. In this work we address the challenging issues of generalization and sparsity of NEs in the context of SMT. Our approach uses the source NE Recognition (NER) system to generalize the training data by replacing the recognized Named Entities with place-holders, thus allowing a Phrase-Based Statistical Machine Translation (PBMT) system to learn more general patterns. At translation time, the recognized Named Entities are handled through a specifically adapted translation model, which improves the quality of their translation. We add a post-processing step to a standard NER system in order to make it more suitable for integration with SMT and we also learn a prediction model for deciding between options for translating the Named Entities, based on their context and on their impact on the translation of the entire sentence. We show important improvements in terms of BLEU and TER scores already after integration of NER into SMT, but especially after applying the SMT-adapted post-processing step to the NER component.

KEYWORDS: Named Entity Recognition, Statistical Machine Translation.

1 Introduction

The correct handling of Named Entities is not an easy task for Statistical Machine Translation. First, Named Entities — person names, organization names, dates, etc. — create a lot of sparsity in the training data. Second, Named Entities of the same type tend to occur in the same context, and thus, they should be treated in a similar way, but a phrase-based SMT model has limited capacity to learn this purely out of data. Finally, Named Entities can be ambiguous (eg. Bush in George Bush vs. blackcurrent bush), and a wrong NE translation can seriously hurt the final quality of the translation.

We propose a framework for integrating Named Entities within SMT, which tries to address all these issues at the same time. First, we try to generalize occurrences of Named Entities in the training data, by replacing the identified named entities by a small number of typed place-holders (one for each NE type: DATE, ORGANIZATION, ...) in order to reduce the sparsity problem, but still preserving some context for the purpose of SMT (all the dates tend to occur in similar contexts, different from the contexts in which person names occur). This generalization allows us to learn a better translation model, and to re-use the generalized patterns for rare (or unseen) Named Entities, in order to ensure a better translation for these NEs. Second, an external NE-translator (or multiple NE-translators for different NE types) is integrated in this framework, thus ensuring correct NE translation.

Third, we address the problem of adapting the NER system itself specifically for the purpose of improving the SMT task. There are few works reporting significant improvements over a baseline after Named Entities integration[1] (eg. from 8.7 to 13.3 of BLEU for Bangla-English (Pal et al., 2010), from 47 to 48.7 of BLEU for Hindi-English (Huang, 2005)). Others report rather low (sometimes negative) impact of Named Entity integration with SMT (0.3 BLEU gain for French-English in (Bouamor et al., 2012), 0.2 BLEU gain for Arabic-English in (Hermjakob et al., 2008), 1 BLEU loss for Chinese-English in (Agrawal and Singla, 2010)).

This is a disappointing result given how important correct NE translation is for overall translation quality. Possible reasons for this result (some of them identified in (Hermjakob et al., 2008)) include:

- Errors of the Named Entity Recognizer itself;

- The external NE-translator is often blind to the type of the NE; however, different treatments can be necessary for different types (e.g. some entities may require transliteration, others a specific kind of translation, and still others should not be translated);

- Often the integration of Named Entities is done by constraining a phrase-based model to producing a single candidate translation for a NE (as generated by an external NE translator): this may prevent the phrase-based model from using known phrases containing the same NE in a larger context, which might have led to more accurate translations.

We note that standard NER systems are designed for Information Extraction tasks and that the Named Entity structure required for these tasks may be different from that required for SMT. In this work we study how the NE structure may be adapted for integration within SMT and

---

[1]Note that not all works explicitly report the gain due to NE integration, but rather the joint gains due to the multiple factors involved. We only mention works in which the specific impact of the NER component is reported explicitly.

propose a post-processing method for a standard NER system in order to adapt this structure. We also propose a way to restrict the use of an external NE translator to those cases where calling it is really useful for the SMT task. First, we apply a set of general rules in order to make the NE structure more suitable for SMT. Next, we develop a prediction model which is able to choose for each NE which translation model is the best to translate it: either an external NE-translator (possibly chosen between multiple options), or the standard SMT model (in this case no special treatment is done for this NE).

The remainder of this paper is organized as follows. Section 2 describes our approach: we first present the general framework we propose for NER integration within SMT, and we then describe the post-processing and prediction steps for NER, which make NE integration more suitable for SMT. Section 3 presents an overview of the related work. Section 4 describes the experimental results and we conclude in Section 5.

2   Proposed Approach for the NE-enriched SMT model

2.1   Translation architecture

The framework that we propose can be summarized by the steps illustrated in the following example:

Src: *This paper illustrates the actions scheduled in Measure 6.2 " Co-operation in agriculture " of the Programme of the European Initiative Interreg II Italy - Albania, being implemented in Apulia since 1996.*

(1) First, we detect Named Entities in the source sentence and replace them with place-holders defined by the type of the NE (eg. DATE, ORGANIZATION, LOCATION): this gives us two types of objects that we need to translate: *reduced source sentences* (source sentences with place-holders) and original *named entities*;

> **Reduced Src:** *This paper illustrates the actions scheduled in Measure 6.2 " Co-operation in agriculture " of the Programme of the European Initiative Interreg II +NE_LOCORG_COUNTRY - +NE_LOCORG_COUNTRY , being implemented in +NE_LOCORG_CITY since +NE_DATE .*
>
> **NEs:** *Italy[LOCORG_COUNTRY], Albania[LOCORG_COUNTRY], Apulia[LOCORG_CITY], 1996[DATE]*

(2.1) The *reduced translation model* (able to deal with the place-holders) is applied to the reduced source sentence and generates a reduced translation:

> **Reduced Translation:** *cet article illustre les actions prévues dans la mesure 6.2 " la coopération en agriculture " du programme de l' initiative interreg II +NE_LOCORG_COUNTRY - +NE_LOCORG_COUNTRY , mis en oeuvre à +NE_LOCORG_CITY depuis +NE_DATE .*

(2.2) An external NE translator is used for translating the replaced NEs; In principle, multiple NE translators can be used, depending on the nature of the Named Entity: a NE can stay

untranslated or be transliterated (eg. PERSON), or its translation can be based on hand-crafted or automatically learned rules (eg. UNITS, 20°C = 68°F), or on an external Named Entity dictionary (which can be extracted from Wikipedia or from the parallel texts):

NE translation: *Italy=Italie, Albania=Albanie, Apulia=Pouilles, 1996=1996*

(3) Finally, Named Entity translations are re-inserted into the reduced translation (this uses the alignment produced internally by the SMT system for deciding which target place-holder corresponds to each source place-holder).

Complete Translation: *cet article illustre les actions prévues dans la mesure 6.2 " la coopération en agriculture " du programme de l' initiative interreg II Italie - Albanie , mis en oeuvre à Pouilles depuis 1996 .*

This integration of NER into SMT already addresses several problems of NE translation:

- First, assuming that NER is able to detect named entities, the approach avoids wrongly translating a NE as if it were a standard lexical expression;

- Second, the approach can translate NEs differently based on their identified type;

- Third, the reduced translation model is based on a generalization of training data which reduces sparsity, and, as a consequence, is able to learn a better model: the generalized patterns are helpful for dealing with rare or unseen Named Entities (eg. the bi-phrase *on +NE_DATE = le +NE_DATE* can be used to translate any date, and not only those seen in the training data).

2.2   NER adaptation for SMT

A weak point of our architecture is that the identification and processing of NEs is only loosely dependent on the SMT task. To get a tighter integration, we apply a post-processing method to the output of the NER system in order 1) to modify the NE structure for a better fit with SMT, and 2) to choose the NEs that have a potential to improve the final translation. We propose a hybrid post-processing, where:

- first, on each source sentence, a set of post-processing rules is applied to the NER output,

- second, a prediction model is applied to the NER output in order to choose only those Named Entities for specific NE-translation that can actually be helpful for SMT purposes; the prediction model is trained to optimize the final translation evaluation score.

We show empirically the importance of each of these steps in section 4.

2.2.1 Rule-based adaptation of NER systems for SMT purposes

Since numerous high-quality NER systems are ready to use, it is more reasonable to take advantage of them for SMT than to develop a new NER system from scratch. NER systems are usually developed for the purposes of information extraction, where the NEs are inserted in a task-motivated template. This template determines the scope and form of NEs. In the case of SMT the "templates" into which the NEs are inserted are sentences. This means that the NEs should be defined according to sentence-translation oriented criteria, because this ensures better quality of the model acquired from sentences containing place-holders for the named entities. In other words, the place-holders should not introduce a similar sparsity factor into the translation model to what the original NEs did. Thus existing NER systems may need some adaptation for SMT.

We consider the following requirements for designing the scope and the form of the NEs for SMT:

- The NEs extracted should not contain common nouns that might be relevant in an IE system, but do not need special translation: titles (Mr, Vice-President, etc.) and various other common nouns (street, road, number etc.). These elements should be removed from the scope of the NEs for SMT, and should be translated as parts of the reduced sentence, and not in the NE translation system.

- The NEs are embedded in various syntactic structures in the sentences, and often the units labeled as named entities contain structural elements in order to yield semantically meaningful units for IE. These structural elements are useful for training the reduced SMT model, and thus they should not be part of the NE. E.g. *le 1er janvier* should rather produce *DATE(1er janvier)* than *DATE(le 1er janvier)*.

The adaptation is rule-based. Given an existing NER system, the adaptation is executed along the following steps:

1. Extract NEs from a corpus relevant to the domain;

2. Either manually or automatically identify the list of common nouns within the NEs (titles, geographical nouns, etc.);

3. Either manually or automatically identify the list of function words at the beginning of NEs;

4a. If the NER system is a black box:
   - Define rules (e.g. POS tagging, list, pattern matching) to recognize the common nouns and the function words in the output of the NER system;
   - Post-process the NEs extracted so that the common nouns and the function words are deleted;

4b. If the source code of the NER system is available: Modify the source code so that the common nouns and function words do not get extracted.

2.2.2 Machine Learning extension of NER adaptation

The previously defined rules allow us to deal with a segmentation of Named Entities that is more suitable for SMT purposes: e.g. this segmentation may separate clearly the non-translatable units composing a person name from its context (ex: Mr.[context] White[non-translatable unit]). However, the importance of certain NEs or NE types for SMT may vary across different domains and text styles. It may also be dependent on the SMT model itself: simple Named Entities that are frequent in the data on which SMT was trained are already well-translated by a baseline model, while the call for an external NE-translator will make the process more complex, and in some cases, produce worse results (due to the lack of context).

The impact of one specific Named Entity on the final translation quality may depend on different factors: NE context, NE frequency in the training data, the type of the NE, the reliability of NE-translator, etc. The impact of each of these factors may be heterogeneous across different domains and styles of the text, and a rule-based approach is not suitable to address this problem in its generality.

We propose to learn a *prediction model*, based on the features that control these different aspects, which will be able to predict the impact that the special treatment of a specific Named Entity could potentially have on the final translation. The main objective of this model is to select only NEs that can improve the final translation, and reject the NEs that can hurt or make no difference for the final translation. In order to achieve this objective, we create an appropriate training set as described below.

In what follows we refer to the baseline SMT model as SMT and to the NE-enriched SMT model as SMT_{NE}. For the prediction training we create a labelled training set out of a set of parallel sentences $(s_i, t_i), i = 1..N$.

- For each $i = 1..N$:
    - translate s_i with the baseline SMT model: $SMT(s_i)$;
    - For each NE ne_k found by NER (and post-processed by a rule-based step) in s_i:
        * translate $s_i|_{ne_k}$ with the NER enriched SMT model: $SMT_{NE}(s_i|_{ne_k})$; ne_k is replaced by a place-holder in s_i, and external NE-translator is used to translate ne_k;
        * compare $SMT(s_i)$ and $SMT_{NE}(s_i|_{ne_k})$ by comparing them to the reference translation t_i: we denote the corresponding evaluation scores by $score(SMT_{NE}(s_i|_{ne_k}))$, $score(SMT(s_i))$ (we may use any standard MT evaluation metric, suitable for sentence-level evaluation);
        * the label of the named entity ne_k is based on the comparison between $score(SMT_{NE}(s_i|_{ne_k}))$ and $score(SMT(s_i))$: positive if $score(SMT_{NE}(s_i|_{ne_k})) > score(SMT(s_i))$ (meaning that NE-enriched system produces a better translation than a baseline), and negative otherwise.

A classification model trained on a training set created in this way will be optimized (by construction) to choose the NEs that improve the final translation quality; the features for this classification model are detailed in section 4.2.2.

This method can also be extended for the case where multiple NE translation systems are available: eg. do not translate/transliterate (person names), rule-based (eg. UNITS, $20°C =$

68°F), dictionary based, etc. In this case the translation prediction model can be transformed into a multi-class labelling problem, where each class corresponds to the model that should be chosen for a particular NE translation model (including the model that do nothing and let baseline the SMT model to deal with NE translation).

2.3 Training NE-enriched SMT

To apply he translation framework described above, first, we need to train a reduced translation model that is capable of dealing with the place-holders correctly. The training of the reduced translation model requires a reduced parallel corpus (a corpus with both source and target Named Entities replaced with place-holders). In order to keep consistency between source and target Named Entities we project the source Named Entities to the target part of the corpus using the statistical word-alignment model (obtained with GIZA++, similar to (Huang and Vogel, 2002)).

Next, we train a phrase-based statistical translation model on the corpus obtained in this way, which allows us to learn generalized patterns (eg. *on +NE_DATE = le +NE_DATE)* for better NE treatment. The replaced Named Entity and its projection are stored separately in a Named Entity dictionary that can then be re-used for NE translation.

When every source Named Entity that was correctly projected to the target sentence is systematically replaced by a place-holder, the translation model trained on such a corpus will not be able to translate the original NEs (they will never or very rarely occur in the resulting training data). This is in contradiction with our prediction model, which may choose to replace or not a NE with a place-holder depending on its context, requiring the ability to translate both a reduced and a non-reduced sentence.

In order to meet this requirement we train a hybrid NE-enriched model, which replaces a NE by a place-holder with probability α: a model trained on a corpus created in this way will indeed be able to translate the frequent NEs in their original form, but at the same time it allows generalization (which is especially important for rare NEs). This hybrid model was inspired by (Bisazza and Federico, 2012), where a hybrid LM was trained in a similar way.[2]

Possible models for the NE-translator include:

- NEs extracted out of parallel corpora by projection of source NEs on the target side can be re-used as NE-translations at the translation step;

- another option is to create an adapted SMT model for NE translation: perform tuning of the baseline PBMT on the subset of extracted NEs (such a model can be useful for the Named Entities that should be translated, but are not directly available in the NE dictionary, eg. *General Division of Land Management, Housing and Patrimony* [ORGANIZATION])

3 Related Work

The mainstream approach for Named Entity integration into an SMT framework is to detect a NE (with an existing NER) and apply an external translation model (NE-translator) to translate the detected NE. The translation proposed by the external model is then integrated into SMT

---

[2]In our experiments, we take $\alpha = 0.5$.

a) as a default translation (Li et al., 2009; Huang and Vogel, 2002), b) added dynamically to the phrase-based table to compete with other phrases (Turchi et al., 2012; Hermjakob et al., 2008; Bouamor et al., 2012), c) replaced by a fake (non-translatable) value, which is replaced by the initial Named Entity once the translation is done (applied for non-translatable NE in (Tinsley et al., 2012)).

This approach mainly addresses the disambiguation issue when translating Named Entities (given that NER is actually able to disambiguate properly), in order to guarantee a correct NE translation.

The sparsity problem is partially addressed either by extracting bilingual Named Entities from the parallel corpus and appending them to the training data, in order to improve the alignment procedure (Bouamor et al., 2012; Okita et al., 2010). However, this approach does not allow to generalize the information learned from the training data for new, unseen Named Entities.

Several "soft" integrations of the NE-translator were previously suggested (Turchi et al., 2012; Hermjakob et al., 2008; Bouamor et al., 2012), where a translation proposed by the NE-translator competes with other phrases of the phrase-table. This allows not to decrease the final translation quality when a wrong NE is proposed by the NER system (either because it is not suitable for an external NE-translator, or because of an error has been done by NER). But this approach does not allow to correct the output of the NER system, and at best it allows simply not to decrease the translation quality due to a wrongly-formed Named Entity, but there is no possibility to improve the final translation in this approach.

The closest work to ours is the one by (Hermjakob et al., 2008), who addresses a problem of NE transliteration for Arabic-English translation. Similar to our approach, the authors propose to adapt the transliteration model for the translation task, and to "learn" when the transliteration is actually helpful for SMT, rather than trust blindly the NER and transliterate every output of the NER system (which may often introduce new errors). However, the way this adaptation is done is very different from what we propose. It relies on annotations done on the parallel training corpus, where each Arabic token/phrase is marked if its transliteration is found in the corresponding English sentence. This annotated corpus is then used to train a transliteration model. However it is not straightforward that the thus learned transliteration model is actually one that improves the final translation quality: the authors report similar results in terms of BLEU to those of a baseline SMT, although the model appears to improve the Named Entities translation (measured in terms of NEWA (Hermjakob et al., 2008)). This indicates that although overall NE translations were improved, probably the context in which they occurred was less accurate, or in some cases the errors done by NER (or the transliterator) led to worse translation. Our NER postprocessing approach optimizes explicitly the final translation score, and can actually be complementary to the approach taken by (Hermjakob et al., 2008). Moreover, some heuristics used by (Hermjakob et al., 2008) (such as applying the transliteration model only to NEs that occurred less than 50 times in training data) can be taken into account in our approach in a more flexible way, at the same time as other important features (e.g. the context in which NE occurs, the confidence of the proposed transliteration etc.).

Table 1: Statistics for the train and test data.

Data set	Nb units	Nb tokens En	Nb tokens Fr
train	152525	3176875	2914542
extra monolingual data	118946	-	4331604
dev-set, MERT-tuning	1100	36484	40474
dev-set, NE prediction mode	1100	36672	41052
test-abstracts	426	45115	58549
test-titles	2000	23888	30786

4 Experiments

4.1 Data and baseline

The training set of parallel sentences was further extended with a subset of the JRC-Aquis[3] corpus, based on the domain-related Eurovoc categories. Overall, the in-domain training data consist of 3M tokens per language.

We have extracted two development sets containing both abstracts and titles. The first dev-set was used for the MERT optimisation of the NE-reduced translation model. The second dev-set was used for training the NE prediction model2.2.2. Both dev-sets were extracted from truly in-domain data (INRA & FAO)

We tested our approach on two different types of texts extracted from in-domain data: 2000 titles (test-titles) and 500 abstracts (test-abstracts). Statistics about the train and test data are given in table 1.

We used a phrase-based SMT model trained by Moses(Koehn et al., 2007) with standard Moses settings (5-gramm LM, lexicalized reordering) on this data as the baseline translation system for our experiments.

4.2 NER adaptation

4.2.1 Rule-based NER adaptation

As a baseline NER system we used the NER component of the Xerox Incremental Parser (XIP (Aït-Mokhtar et al., 2002)) for English. The baseline NER system is rule-based and recognizes a large number of different Named Entities: date, person, numerical expressions, location names, organization names, events.

We ran XIP on a development corpus and extracted lists of NEs: PERSON, ORGANISATION, LOCATION, DATE. We then identified a list of common names and function words that should be eliminated from the NEs. In the XIP grammar NEs are extracted by local grammar rules as groups of labels that are the POS categories of the terminal lexical nodes in the parse tree. The post-processing consisted in re-writing the original groups of labels by ones that exclude the unnecessary common nouns and function words (see section 2.2.1).

---

[3]http://langtech.jrc.it/JRC-Acquis.html

4.2.2 Prediction model for choosing NE translation model

The prediction model for SMT adaptation relies on the following features:

- Named Entity frequency in the training data;

- confidence in the translation of NE dictionary; (if ne_s: source named entity, ne_t: translation suggested for ne_s by NE dictionary, we measure confidence as $p(ne_t|ne_s)$ estimated on the training data used to create NE dictionary);

- a collection of features defined by the context of the Named Entity: the number of features in this collection corresponds to the number of trigrams that occur in the training data of the following type: a named entity place-holder extended with its 1-word left and right context (eg. *the +NE_DATE,*);

- the probability of the Named Entity in the context, estimated from the source corpus (3-gram Language Model);

- the probability of the place-holder replacing a Named Entity in the context (3-gram reduced Language Model);

The corpus used to train the prediction model contains 2000 sentences (a mixture of titles and abstracts). A labelled training set is created out of a parallel set as described in 2.2.2. We used the TER (translation edit rate) score for measuring individual sentence scores. Overall, we obtain 461 labelled samples, with 172 positive examples, 183 negative examples, and 106 neutral examples (the samples where both SMT_{NE} and SMT provide the same translation). We learn a 3-class SVM prediction model and we choose to replace only the NEs that are classified as positive at test time.

4.3 NE-enriched SMT training

We train a hybrid reduced translation model replacing a Named Entity by a place-holder with probability $\alpha = 0.5$ as described in section 2.3. The NE-translator performs as follows:

- First, it checks whether a NE translation is available in the NE dictionary extracted from the parallel corpus (which contains 11347 entries);

- If no translation is found in the NE dictionary, a baseline SMT model, with weights tuned on a subset of NEs extracted from the parallel corpus, is used as a back-off.

4.4 Evaluation

We evaluate the performance of different translation models using both BLEU (Papineni et al., 2001) and TER (Snover et al., 2006) metrics. We compare the following translation models:

- SMT: a baseline phrase-based statistical translation model without Named Entity treatment;

- $SMT_{NE-baseline}$: NE-enriched SMT (described in 2.1) where the baseline NER is used (no NER post-processing is done);

- RB-adapted SMT_{NE-RB} : SMT_{NE} where only the first post-processing step (rule-based NE structure modification described in 2.2.1) is applied to the baseline NER;

- SMT_{NE-ML}: SMT_{NE} where only the second post-processing step (prediction model described in 2.2.2) is applied to the baseline NER;

- $SMT_{NE-full}$: SMT_{NE} relying both on rule-based and machine learning post-processing steps for NER.

We also compare the results of our NE-enriched system to the approach used by (Turchi et al., 2012) where the NE translations provided by an external dictionary (the NE dictionary extracted from the parallel corpus in our case) are suggested as dynamic bi-phrases (using Moses XML tagging mechanism) to the decoder. We refer to the approach used in (Turchi et al., 2012) as $SMT_{NE-Turchi}$; this is the soft NE integration into the model (soft XML tagging option of Moses), which may choose the NE translation between the one suggested by the NE dictionary and the one suggested by the baseline SMT during the decoding process. In principle, this NE integration is more flexible than the pipeline approach we adopt. However, this approach does not have the generalization capability of our NE-enriched model.

Table 2: Results for NER adaptation for SMT

Model	test-titles		test-abstracts	
	BLEU	TER	BLEU	TER
SMT (baseline)	0.3135	0.6566	0.1148	0.8935
$SMT_{NE-Turchi}$	0.3135	0.6565	0.1149	0.8934
$SMT_{NE-baseline}$	0.3213	0.6636	0.1211	0.9064
SMT_{NE-RB}	0.3258	0.6605	0.1257	0.8968
SMT_{NE-ML}	0.3371	0.6523	0.1228	0.9050
$SMT_{NE-full}$	**0.3421**	**0.6443**	**0.1341**	**0.8935**

The translation results for the models described above are reported in Table 2.

First, we abstract from NER adaptation, and compare two approaches relying on the non-adapted NER, to evaluate our NE-enriched SMT model. We show that our approach $SMT_{NE-baseline}$ performs better than $SMT_{NE-Turchi}$. We believe that this gain is due to the generalization capacity of our model. Indeed, since the training data we used is relatively small, the sparsity issue is very important in this setting, and the capacity to generalize the observed NE occurrences helps our model. We see that $SMT_{NE-Turchi}$ performance is very close to the baseline SMT. This is probably due to the fact that the only NEs that are integrated are those that are already present in the training corpus, and no external knowledge was injected. This is, however, also the case for our model, and we believe that adding an external NE dictionary might improve both models.

Second, we note that each of the NER adaptation post-processing steps (SMT_{NE-RB} and SMT_{NE-ML}) brings improvements compared to the case when non-adapted NER is used ($SMT_{NE-baseline}$). Finally, we see that the combination of both steps gives the best results, which proves that these two steps complement each other and are both important for the final translation quality.

Table 3: Named Entity density in the test data

NE type	test-titles		test-abstracts	
	NEs detected	NEs selected	NEs detected	NEs selected
DATE	191	48 (25%)	121	32 (26%)
LOCATION	127	28 (22%)	61	20 (32%)
LOCORG	614	190 (30%)	189	44 (23%)
ORGANISATION	132	38 (28%)	210	33 (15%)
PERSON	95	44 (46%)	79	31 (39%)
EVENT	3	1 (33%)	3	0
UNIT	6	0	82	3 (3%)
PERCENT	2	1 (50%)	84	20 (23%)
Total	1170	350 (29%)	823	183 (22%)

4.4.1 Error analysis

We have performed some error analysis to find out the interaction between various aspects of our model with the final translation performance.

First, we carried out a small-scale manual evaluation of NER over around 500 entities for English. The recall of for all the NEs (including non-detectable NE types) was 53% and the precision was 86%. The types of NEs not detected but potentially relevant were projects, titles and biological entities. The worse performance among detectable NE types was observed for the organization names (precision: 80%, recall: 68%). This performance can be explained by the domain specificity of our data which is very different from the one (news articles) which was used for NER development.

Second, we looked at the NE density in the corpus and how the integration of the prediction model impacts it. Table 3 reports the number of different NEs (by type) detected in total in each test set, and the number of NEs that were selected by the prediction model (meaning that the integration of these NEs has the potential to improve the final translation). First, we see that we select only 29% of the total entities detected in the titles test set, and even fewer (22%) in the abstracts test set. We also observe that NEs density is lower in the abstracts than in the titles, and that the frequency of NE types differs between titles and abstracts: abstracts contain more UNIT and PERCENT types, which are less ambiguous and easier to handle for the baseline SMT. The above mentioned points may also explain lower impact after NE integration on the abstracts test.

We see that the NEs most frequently retained by the prediction model are the PERSON names which are probably the most sparse entities, which can be translated independently of the context. We also see that we retain much fewer ORGANISATION types in the abstracts test compared to the titles test: this is due to the fact that the organization names that occur in the titles are frequent acronyms (eg. FAO, ONU, INRA) which are well handled by NER, while abstracts contain more ambiguous and difficult to detect organization names (eg. table 4, ex.3: Confederation of Agricultural Workers).

Finally, table 4 shows some examples extracted from each of the tests on how NE integration impacts the final translation. We see some cases where it is important to have a separate

Table 4: Examples of English-French translations with and without NE integration.

test-titles
1	**Src:** Comparison of the **morphometric indexes** of the grasshopper tippet Schistocerca **gregaria Forskael**, 1775 at **Adrar** and at Tamanrasset (Sahara, Algeria) in 1995 **Baseline:** Comparaison **de l'étude de l'index** grasshopper tippet Schistocerca **gregaria, Forskael** 1775 sur **tomate** et à Tamanrasset (Sahara algérien) en 1995 **NE-full**: Comparaison **des indices morphometriques** de la grasshopper tippet Schistocerca **gregaria Forskael**, 1775 sur le terrain à **Adrar** et à Tamanrasset (Sahara algérien) en 1995
2	**Src:** Decisions in favour of the future generations. Proceedings of the Conference, Brussels, **8 May 1996** [with contributions of George, S.; Rahman A.; Alders, H.; Platteau, J.P.] **Baseline:** Les décisions en faveur des générations futures. Compte rendu de la conférence, bruxelles, **8 peut 1996** [avec les apports de George, S.; Rahman A.; l'aulne, H.; Platteau, J.P.] **NE-full**: Les décisions en faveur des générations futures. Compte rendu de la conférence, bruxelles, **le 8 mai 1996** [avec les apports de George, S.; Rahman A.; l'aulne, H.; Platteau, J.P.]

test-abstracts
3	**Src:** The Author , F. Mellozzini , carries out an in - depth analysis of the objectives of agricultural policy which have arisen during a meeting on " Which kind of agriculture for the 1980 's? " held in Rome by the **Confederation of Agricultural Workers on 18 - 19 October** . **Baseline:** L'auteur, F., Mellozzini exerce une analyse des objectifs de la politique agricole qui ont ainsi présentée au cours de la réunion, sur " dont la nature de l'agriculture de la 1980 ? " tenue à Rome par la **mobilité des salariés agricoles sur 18 - 19 octobre**. **NE-full**: L'auteur, **F. Mellozzini**, exerce une analyse approfondie des objectifs de la politique agricole qui ont ainsi présentée au cours de la réunion sur " qui la nature de l'agriculture pour 1980 ? " tenue à Rome par la **confédération des travailleurs agricoles en 18 - 19 octobre**.
5	**Src:** These studies allowed the drawing up of a balance of its qualities and limits observed , its effectiveness in natural conditions and provide the opportunity to share some ideas on the use **in Africa of the South American auxiliary** . **Baseline:** ces études ont permis l'établissement d'un bilan de ses qualités et limites observés, son efficacité en conditions naturelles et prévoir la possibilité d'action des idées sur l'utilisation **en Afrique du Sud auxiliaires américaine.** **NE-full**:Ces études ont permis l'établissement d'un bilan de ses qualités et limites observés, son efficacité en conditions naturelles et de prévoir la possibilité à part quelques idées sur l'utilisation en **Afrique des auxiliaires d'Amérique du Sud.**
3	**Src:** Farmers are willing to pay between **13.5 percent and 14.5 percent** of the value of the premium rate demanded by insurance companies . **Baseline:** les agriculteurs sont prêts à payer **pour cent entre 13.5 et 14.5 pour cent** de la valeur de la prime taux exigées par les sociétés d'assurance. **NE-full**: les agriculteurs sont prêts à payer **entre 13,5 pour cent et 14,5 pour cent** de la valeur de la prime taux demandées par les compagnies d'assurance.

translation model for the NE itself (ex. 1, 2, 3 and 5). At the same time, we see that although the NE translation did not change, the surrounding context was better translated: ex. 3 "*sur DATE*" vs "*en DATE*", ex.4 : *auxiliary* was better placed in the translation.

Finally, we would like to note that our test set is rather difficult both for NER, and for MT translation. We believe that application of the same NER integration on an easier data set (with higher NER performance) may lead to higher improvements.

5   Discussion and Perspectives

In this work we have addressed the main problems of Named Entities integration into an SMT framework. We have proposed an approach that is able to generalize the Named Entity context observed in the training data and re-use it for new (unseen) NE translations. Our approach can also integrate one or several external NE-translators, and allows to choose an adapted NE-translator for each NE. The choice of the adapted NE-translator model is done via a prediction model that relies on features specific to the NE itself, the context in which it occurs and also the baseline SMT model which is enriched with NER. In addition, we propose a set of NER post-processing rules that allow to modify the NE structure in order to produce better NE segmentation for integration within SMT. We have shown empirically that each aspect of our model is important, and that the combination of all of them leads to the best results (2-3 BLEU points improvement over a baseline for two different test sets).

This framework opens several possible future research directions. First, NER-SMT integration pipeline can be replaced by a confusion network representation, where the best NE translation model will be chosen internally by the decoder. The prediction model scores can serve a basis for assigning a score for each alternative path in the confusion network.

Second, the procedure of creating an annotated training set for learning the prediction model which optimizes the MT evaluation score (described at 2.2.2) can be applied to other tasks than NER adaptation. More generally it can be applied to any pre-processing step done before translation (eg. spell-checking, sentence simplification, reordering, or any other source modification which might help to produce a better translation). The advantage of applying a prediction model to these steps is to make the pre-processing model more flexible and better adapted to the SMT task it is applied to.

Finally, in our experiments we have only used three options for the NE-translator: a NE dictionary extracted out of parallel data, a SMT model tuned for NE translation and a baseline SMT model. There are many other options that need to be explored, among them integrating an external NE dictionary mined from Wikipedia or LinkedData or creating specific translation models for each NE type.

Acknowledgements

References

Agrawal, N. and Singla, A. (2010). Using named entity recognition to improve machine translation. Technical report, Standford University, Natural Language Processing.

Aït-Mokhtar, S., Chanod, J.-P., and Roux, C. (2002). Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(3):121–144.

Bisazza, A. and Federico, M. (2012). Cutting the long tail: Hybrid language models for translation style adaptation. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, pages 439–448.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012). Identifying multi-word expressions in statistical machine translation. In *In LREC 2012, Seventh International Conference on Language Resources and Evaluation*.

Hermjakob, U., Knight, K., and Daumé III, H. (2008). Name translation in statistical machine translation learning when to transliterate. In *Proceedings of ACL-08:HLT*.

Huang, F. (2005). *Multilingual Named Entity extraction and translation from text and speech*. PhD thesis, Language Technology Institute, School of Computer Science, Carnegie Mellon University.

Huang, F. and Vogel, S. (2002). Improved named entity translation and bilingual named entity extraction. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI '02, pages 253–, Washington, DC, USA. IEEE Computer Society.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Li, M., Zhang, J., Zhou, Y., and Chengqing, Z. (2009). The CASIA statistical machine translation system for iwslt 2009. In *Proceedings of IWSLT 2009*.

Okita, T., Maldonado Guerra, A., Graham, Y., and Way, A. (2010). Multi-word expression-sensitive word alignment. In *Proceedings of the 4th Workshop on Cross Lingual Information Access*, pages 26–34, Beijing, China. Coling 2010 Organizing Committee.

Pal, S., Kimar Naskar, S., Pecina, P., Bandyopadhyay, S., and Way, A. (2010). Handling named entities and compound verbs in phrase-based statistical machine translation. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2001). Bleu: a method for automatic evaluation of machine translation.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Tinsley, J., Ceausu, A., and Zhang, J. (2012). PLUTO: automated solutions for patent translation. In *EACL Joint Workshop on Exploitng Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra): Proceedings of the workshop, EACL 2012*.

Turchi, M., Atkinson, M., Wilcox, A., Crawley, B., Bucci, S., Steinberger, R., and Van der Goot, E. (2012). ONTS: "Optima" news translation system. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*.