Machine Translation of Speech-Like Texts: Strategies for the Inclusion of Context

Rachel Bawden
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay
rachel.bawden@limsi.fr

ABSTRACT

Whilst the focus of Machine Translation (MT) has for a long time been the translation of planned, written texts, more and more research is being dedicated to translating speech-like texts (informal or spontaneous discourse or dialogue). To achieve high quality and natural translation of speech-like texts, the integration of context is needed, whether it is extra-linguistic (speaker identity, the interaction between speaker and interlocutor) or linguistic (coreference and stylistic phenomena linked to the spontaneous and informal nature of the texts). However, the integration of contextual information in MT systems remains limited in most current systems. In this paper, we present and critique three experiments for the integration of context into a MT system, each focusing on a different type of context and exploiting a different method: adaptation to speaker gender, cross-lingual pronoun prediction and the generation of tag questions from French into English.

RÉSUMÉ

Traduction automatique de l'« oral-écrit » : Stratégies pour l'intégration du contexte

Bien que la Traduction Automatique (TA) se soit concentrée jusqu'à présent sur la traduction de textes écrits et édités, de plus en plus de travaux sont consacrés à la traduction de textes informels et spontanés (discours et dialogues). Pour traduire de tels textes relevant de l'« oral-écrit », il devient indispensable de prendre en compte des informations contextuelles, qu'elles soient de nature extra-linguistique (identité du locuteur, interaction entre le locuteur et l'interlocuteur) ou linguistique (coréférence et phénomènes stylistiques propres à la parole). Or l'intégration d'informations contextuelles dans les systèmes de TA reste limitée dans la plupart des systèmes actuels. Dans cet article, nous présentons et analysons trois expériences d'intégration du contexte dans un système de TA mettant en jeu des formes de contexte et donc des méthodologies différentes: l'adaptation au genre du locuteur, la traduction de pronoms et la génération de « tag questions » anglaises à partir du français.

MOTS-CLÉS : traduction automatique, contexte, parole, genre, pronoms, tag questions.

KEYWORDS: machine translation, context, speech-like texts, gender, pronouns, tag questions.

1 Introduction

Speech-like texts (social networking, speech transcriptions, subtitles and other informal written exchanges)[1] pose new challenges for Natural Language Processing (NLP) and in particular Machine

[1]Our focus is on the speech-like nature of texts, rather than the medium of communication (written versus oral). The processing of oral discourse is an important research topic, but is not the study of this paper. We therefore choose to refer to the genre of informal, spontaneous productions as being speech-like, whether they are transcriptions of oral discourse or were

Translation (MT), when compared to the genres of parliamentary discourse and journalistic texts, which have been the main focus of text processing applications to date. As the need for high quality MT of this more informal genre increases, more and more MT research is being dedicated to exploring the difficulties speech-like texts present and how we might hope to overcome them (Hardmeier, 2012; Guillou, 2016).

Speech-like texts are fundamentally different from planned written texts for a number of different reasons. One of these is their highly contextualised nature. Whilst it is true that all texts are related to a certain genre and context of production, the extra-linguistic context of a speech-like production is very often essential to understanding its linguistic content and therefore to producing accurate and coherent translations. Unlike journalistic texts, which are addressed to a heterogenous audience, absent at the moment of the text's production, in discourse, the speaker and the interaction with the listener can greatly influence the style (in terms of formality and politeness), vocabulary choices and even the grammatical agreement (in languages that have grammatical gender agreement for example) of the texts. Spontaneity of production allows speakers to continuously adapt to each other's reactions, making the aforementioned contextual aspects even more important. There is also the added possibility that speakers align to each other's linguistic choices, make reference to entities mentioned in the other speaker's speech turn and rely more on implicit common ground to root their speech, potentially resulting in more ambiguity for automatic processing.

Traditional MT methods suffer from a considerable flaw when it comes to dealing with context, this being that sentences are processed independently of each other, and translation choices are often made based on very local context, which is especially true of phrase-based statistical MT (SMT) techniques. Whilst attentional Neural MT (NMT) approaches can partially alleviate this problem, the degree to which we can control which context is used in translation is limited. As for extra-linguistic information, it must first be identified and then integrated into the translation process, through pre-processing, during decoding or as a post-edition step.

We illustrate the different strategies that can be used to integrate contextual information into the MT process by presenting a discussion orientated around three separate experiments for the language pair English-French, each illustrating the integration of a different type of contextual information. The scope and nature of context are very different in each case, presenting different challenges to MT. The experiments described illustrate the nature of the different problems faced, as well as provided a basis for the critique of the methods used, highlighting the fact that the problem is far from being solved.

The first experiment is an illustration of the use of domain adaptation, a very simple technique used to take into account a very coarse-grained and static type of information, that of speaker gender, which applies on the sentence-level (Section 2). The second experiment deals with a finer notion of context, necessitating a more sophisticated modelling of an utterance's linguistic context, in a task to predict the French translation of the English subject pronouns *it* and *they*, which, unlike their equivalent French pronouns, are unmarked for gender (Section 3). In this case, a very specific contextual element (the antecedent) determines the translation of a very specific textual element (the pronoun). Finally, in Section 4, we present our third experiment, a post-editing task to improve the translation of a particular stylistic phenomenon common in spoken English, the English tag question, whose usage is largely determined by speaker attitude and the interaction between speaker and listener. Whilst the phenomenon affected by context (the presence and form of the English tag question) is clearly identified, determining which context can be useful is much more complex.

---

never produced orally, as in the case of social network posts and discussions.

2 Integrating speaker gender through domain adaptation

Speaker identity (as defined by demographic factors such as age, gender and social background, as well as other aspects such as personality and mood) greatly influences the way in which speakers communicate, in terms of lexical choices, syntactic structure, politeness strategies, etc. This variation has long been studied in the sociolinguistics literature, in particular with respect to the impact of gender on communication style (Lakoff, 1975; Coates, 1986). Whilst speaker adaptation has long been a part of speech recognition systems (Kuhn *et al.*, 1998), only recently has research in NLP turned towards adapting systems to take into account this extra-linguistic information (cf. works by Volkova *et al.* (2013) for sentiment analysis and Hovy (2015) for three separate NLP classification tasks). Speaker identity is just as important in MT; van der Wees *et al.* (2016) show how MT performance fluctuates with respect to individual speaker identity, gender and register in five different language pairs for film and TV subtitles, and Mirkin *et al.* (2015) find that MT translated output does not conserve speaker personality and gender traits as well as human translations do.

One of the simplest methods for integrating coarse-grained, sentence-level information such as aspects of speaker identity is to perform domain adaptation. *Domain* in this context refers to a broader concept than the genre of text and can include any sentence-level aspect representing the class of sentence, such as sentence type, speaker gender and formality. Domain adaptation is a well-known approach in MT and is used almost systematically. In its simplest form, it implies simply choosing an in-domain development set to tune the system (cf. Pecina *et al.*, 2012 for a study of the effectiveness of in-domain tuning for SMT systems when dealing with a low-resource domain). But its implementation can be more complex, for example through similarity-based data selection techniques for both training and tuning data (Axelrod *et al.*, 2011) or by weighting different components of the MT system that have been trained on class-partitioned data (Foster & Kuhn, 2007; Finch *et al.*, 2009).

2.1 Experiments and results

Our preliminary experiments into the use of domain adaptation to take into account the gender of the speaker (Bawden *et al.*, 2016) give us the opportunity to critique this method, in particular in terms of the possibility of extending the technique to multiple aspects of speaker identity. Similar experiments on gender adaptation were performed concurrently by van der Wees *et al.* (2016) and Wang *et al.* (2016). Speaker gender is a good starting point for integrating information concerning speaker identity, since it is relatively easy to identify (through knowledge of the person's name if the text is written) and is generally binary. As well as influencing word choice and style, speaker gender is also explicitly visible in languages which display grammatical gender agreement with the speaker. For example, the English sentence "I am happy" would be translated into French "Je suis heureuse" for a female speaker and "Je suis heureux" for a male speaker. When translating from English into French, standard MT systems are limited to selecting the most probable form given the linguistic context, which in most cases is the masculine form "heureux".

We performed a series of simple domain adaptation experiments, consisting of adapting a standard SMT system trained using Moses (Koehn *et al.*, 2007), by modifying both the data used to train the components of the system and the data used for tuning. Our aim was to evaluate the effect on translation performance of training gender-specific language models and gender-specific translation models and of tuning to a gender-specific dataset. The idea of domain adaptation by data partitioning is to annotate the sentences of the data according to the different aspects being studied (the different

domains in the broader sense of the term) and to use the data selectively in the training and tuning of the MT system based on these annotations. In our case, the domain was the gender of the speaker, so each dataset was divided into two, corresponding to the sentences spoken by female characters and those spoken by male characters.[2] We used subtitles annotated for speaker gender, taken from the TVD *Big Bang Theory* reproducible corpus (Roy *et al.*, 2014), resulting in approximately 10,000 gender-annotated utterances. The results of the automatic evaluation are summarised in Table 1, using two automatic metrics, BLEU (Papineni *et al.*, 2002) and METEOR (Lavie & Denkowski, 2009).[3] The baseline system is shown in the first row of the table – no specific adaptation is performed and all development data (containing both male and female speakers' utterances) is used to tune the model.

		BBT-test$_{\text{male}}$		BBT-test$_{\text{female}}$	
Model adaptation	Tuning data	BLEU	METEOR	BLEU	METEOR
(i) Choice of the tuning set					
∅	all	23.91	0.434	25.16	0.450
∅	male	**24.09**	**0.438**	**25.72**	0.450
∅	female	23.67	0.431	25.22	0.446
(ii) Addition of a gender-specific language model					
+LM$_{\text{male}}$	all	**24.17**	**0.436**	24.80	0.447
+LM$_{\text{female}}$	all	23.35	0.430	24.13	0.443
+LM$_{\text{male}}$	male	23.92	0.435	25.39	0.448
+LM$_{\text{female}}$	female	23.97	0.444	**26.25**	**0.459**
(iv) Addition of a gender-specific language model and translation model					
+LM$_{\text{male}}$+TM$_{\text{male}}$	all	24.06	0.434	25.36	0.449
+LM$_{\text{female}}$+TM$_{\text{female}}$	all	23.60	0.431	24.55	0.444
+LM$_{\text{male}}$+TM$_{\text{male}}$	male	**24.18**	**0.436**	**25.69**	**0.451**
+LM$_{\text{female}}$+TM$_{\text{female}}$	female	22.64	0.422	24.91	0.441

Table 1: Translation performance after adaptation of the phrase-based SMT model to gender. LM$_x$ refers to the addition of a language model trained on data labelled as x. Similarly TM$_x$ refers to the addition of a translation model (phase table) trained on data labelled as x.

The results show that small gains in translation performance can be obtained from domain adaptation, although these improvements are not significative. Providing a gender-specific tuning set gives gains of +0.1 BLEU and +0.004 METEOR for the male speakers' utterances, and gains of +0.56 BLEU for the female speakers' utterances. Adding a language model that has been trained exclusively on female or male utterances also provides some very slight improvements for both genders, as does the addition of a gender-specific translation model. However a manual evaluation revealed that the slight improvements seen were not due to improvement in grammatical gender agreement at all. Most improvements were thanks to an improved lexical choice when compared to the baseline translation, followed by lexical additions or deletions. It appears that domain adaptation did indeed adapt slightly to the differently gendered datasets, but not necessarily for the reasons we might think; each dataset has its own distinct lexical properties, and it is possible that the improvements were simply due to minor lexical specificities of the two datasets and not necessarily due to a real gender bias. Data sparsity, in particular for female speakers is a problem with this method, which could in part explain why the male-adapted model performs highly on the female-specific test set.

[2]Male utterances were 3.8 times more frequent than female utterances.

[3]BLEU scores are from 0 to 100, with 100 being the highest score, and METEOR are from 0 to 1, 1 being the highest score.

2.2 Analysis and perspectives

The main problems found with this method is that data partitioning leads to smaller datasets being used for training and tuning. MT is a domain in which having large amounts of data is a must to produce a robust and high-performing model, so the decision to reduce the size of the data used is not one that should be taken lightly.

It is difficult to imagine this method being extended to integrate other coarse-grained, sentence-level features. In theory, the method is simple, requiring simply annotating data according to the set of features (e.g. sentence type, formality, topic). However, in practice, partitioning of the data depending on the set of class labels, whose number is multiplied each time a new feature is added, would result in smaller and smaller datasets on which to train and tune models, which would inevitably lead to degradation in translation performance. A solution to this problem proposed by Saluja *et al.* (2011), which involves weighting different sentences according to a similarity measure between the test sentence and each training sentence (based on their set of class labels), showed some improvements, but in a limited domain. The use of factored translation models (Koehn & Hoang, 2007) allows more linguistic input to be provided for each of the words, but has had limited impact on translation quality and suffers from computational problems when scaling up to large datasets.

Recently, a new method of performing domain adaptation has been introduced, in the context of neural machine translation (NMT). Sennrich *et al.* (2016) show how it is possible to control for politeness on the sentence-level by introducing an additional feature (*side constraint*) as an arbitrary feature at the beginning of each sentence. The flexibility of this approach appears promising for integrating a range of different sentence-level aspects, including extra-linguistic information related to speaker identity.

3 Linguistically motivated cross-lingual pronoun prediction

Other forms of context necessarily require a more sophisticated representation of the linguistic properties of discourse. One of the most studied contextual aspects of discourse is coreference and the translation of pronouns, particularly between languages that do not have the same system of grammatical gender (Le Nagard & Koehn, 2010; Guillou, 2016). Take for example the language pair English-French. French common nouns and pronouns are marked for grammatical gender, whereas the English pronouns *it* and *they* are not. In order to correct translate the pronoun *it* in the English sentence "it was blue", it is necessary to know the grammatical gender of the French coreferential antecedent of the pronoun. For example, if *it* refers to a box (*une boîte*), the correct translation is the feminine pronoun *elle*, whereas if *it* refers to a toe (*un orteil*), the correction translation is the masculine pronoun *il*. There is also a third option (which can be used in certain contexts) to translate *it* using the gender-neutral demonstrative *ce*.

The problem of ensuring accurate pronominal translation has received much interest in the discourse in MT community and a shared task has been organised for the past few years, dedicated to providing solutions to the problem (e.g. Guillou *et al.*, 2016). The task is a cross-lingual pronoun prediction task, for which participating classification systems aim to correctly predict target pronoun forms in the target sentence, based on contextual information in the source and target sentences. We focus on one of the language directions proposed by the 2016 task: English into French, for which the aim was to predict the correct French translation of the English subject pronouns *it* and *they*, from the set of possible classes *il*, *ils*, *elle*, *elles*, *on*, *ce*, *cela* and *OTHER*. An example of the data provided for the

task is shown in Figure 3. Target sentences were provided as lemma-tag pairs, rather than as surface forms, to avoid systems relying solely on the grammatical gender of the immediately surrounded words to determine gender, which is an unrealistic scenario in a real translation setting.[4]

Pronoun class	Lemma+tag	English source sentence	French target sentence	Word alignment
ils	il|PRON	They 're extremely costly .	REPLACE_0 être|VER très|ADV coûteux|ADJ .|.	0-0 1-1 2-2 3-3 4-4

Figure 1: An example of pronoun prediction task data at WMT 2016. The first column represents the class label to be predicted and the second column the lemma and tag of the pronoun. The position of the pronoun to be predicted is marked by the placeholder REPLACE_0 in this example.

A variety of different strategies were used by task participants, who relied on different amounts of contextual, linguistic information. The winning system (Luotolahti *et al.*, 2016) was a stacked recurrent neural network (RNN) system, which did not rely on any other contextual information than the information provided (as shown in Figure 3) and did not look beyond sentence boundaries. A second neural network approach also scored highly in the task: Dabre *et al.* (2016) used a simple RNN architecture with an attention mechanism and trained only using IWSLT data (whereas the winning system was trained on all data provided). Many of the other participants, including the current authors, chose to concentrate on the integration of more linguistic information into simpler (mainly linear) classification systems. Many participants used part-of-speech (PoS)-tagging, parsing and coreference chains. Stymne's (2016) system, a linear classifier using multiple linguistic annotations, was ranked second out of nine systems, showing that simple linear classifiers can perform just as highly as more sophisticated architectures such as neural network systems if good quality linguistic information is used.

3.1 Experiments and results

Our system (Bawden, 2016), followed this second strategy of including as much relevant linguistic knowledge into the system as possible. One of the main aims of our submission was to evaluate the capacity of linguistic tools and resources to provide accurate linguistic knowledge that, if perfect, should be sufficient to predict the correct pronoun. We used a variety of linguistic annotations and heuristics to provide features to a random forest classifier, implemented in Scikit-learn (Pedregosa *et al.*, 2011), relying heavily on linguistic tools and external sources. As source-side information, we used automatic PoS-tagging, dependency parsing and coreference resolution annotations, provided by the Stanford toolkit (Manning *et al.*, 2014). On the target-side, we used dependency parsing using the Mate Parser (Bohnet & Nivre, 2012) and a parse model trained on lemmas and morphological information found in the morphological and syntactic lexicon, the Le*fff* (Sagot, 2010). We also integrated the prediction provided by the language model baseline (Tiedemann, 2016) and identified local windows of syntactic and morphological patterns that were particularly linked to certain classes.[5]

We chose to explicitly perform coreference resolution, using the Stanford toolkit. Since the tool is not available for French, we performed coreference resolution on the English source sentence and

[4] In the 2015 version of the task in which tokens were used on the target side, not a single system beat the baseline system, which was a simple n-gram language-model (Hardmeier *et al.*, 2015). The use of just lemmas in the 2016 task encouraged participants to better model the context and use richer information than just surface forms.

[5] These final features were particularly useful for predicting impersonal *il* pronouns, as well as anomalies in the training data, where the English pronoun was not translated by a French pronoun, and the gold class was therefore the class OTHER.

	Classified as											
	ce	elle	elles	il	ils	cela	on	other	SUM	P (%)	R (%)	F (%)
ce	**54**	1	0	11	0	0	0	2	68	91.53	79.41	85.04
elle	0	**13**	1	6	0	2	0	1	23	41.94	56.52	48.15
elles	1	2	**3**	1	13	1	0	4	25	23.08	12.00	15.79
il	2	7	0	**44**	1	2	1	4	61	61.97	72.13	66.67
ils	0	1	9	0	**56**	0	0	5	71	75.68	78.87	77.24
cela	0	5	0	7	0	**13**	1	5	31	72.22	41.94	53.06
on	0	0	0	0	2	0	**5**	2	9	55.56	55.56	55.56
OTHER	2	2	0	2	2	0	2	**75**	85	76.53	88.24	81.97
SUM	59	31	13	71	74	18	9	98				
Micro-averaged										70.51	70.51	70.51
Macro-averaged										62.31	**60.58**	60.43

Table 2: A decomposition of results on the test set for our submission to the WMT 2016 cross-lingual pronoun prediction task. The results are slightly higher than those reported in the official scoreboard due to the resolution of a minor bug found after the submission deadline.

automatically transferred the information through the alignments to the French target sentence. As shown in Figure 2, (i) we used the automatic alignments to find the pronoun to which the placeholder is aligned, (ii) we then use the English coreference chains to identify the antecedent of the English pronoun, and (iii) we again use the automatic alignment to identify the French antecedent of the pronoun. The gender of the pronoun is given by the French antecedent (found in the lexicon) and number is provided by the number of the English aligned pronoun. Both of these values were used alongside the other features in our classifier.

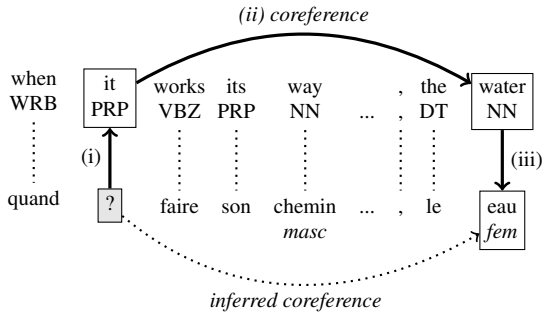

Figure 2: Use of coreference chains to determine gender and number of anaphoric (or, as shown here, cataphoric) pronouns.

The breakdown of the results of our system is shown in Table 2. The official metric for the task was macro-averaged recall, which means that more importance is given to rarer classes, such as *elle*, *elles* and *on*. The system scores ranged from 36.35 to 65.70, with the baseline system at 50.85. The overall score of our system 59.32 put us at sixth position out of nine systems (excluding the baseline system). Our system scored +8.47 points higher than the baseline, showing that exploiting contextual linguistic information is useful for the task.

3.2 Analysis and perspectives

As with all systems submitted to the task, the most difficult classes to predict were the rarest classes: *elle*, *elles* and *on*. In terms of pronouns determined by coreference, where the neural network systems were limited by sentence boundaries, the linguistically-motivated systems were limited by the performance of the tools providing the linguistic annotation. Our analysis of the performance of the Stanford Coreference tool showed that a correct antecedent was provided in only 52.5% of cases and 32% of pronouns were linked only to other pronouns rather than to full noun phrases, even when searching for the noun phrase by transitivity. The tool also fails to identify impersonal uses of pronouns (supplying coreference chains for 18 impersonal pronouns out of 25). Whilst resources and linguistics tools can provide useful information for integration such contextual information, it is important not underestimate their lack of robustness in certain contexts. Neural classification methods, with the use of richer representations of both source and target sentences through embeddings, could provide a good setting for integrating this contextual information more effectively, without having to rely as much on the accuracy of the annotation tools.

4 Stylistic choices: the case of tag questions

The final focus of this paper is on an even more complicated notion of context, which is more difficult to characterise than the two previous types of contexts discussed in the two previous sections. Unlike speaker identity or grammatical gender in coreference chains, which are deterministic and easily discernible, we focus here on one particular aspect of stylistic choice, the English tag question. Common in spoken English (particularly British English), the tag question, which is best known in its canonical form, formed of an auxiliary verb and a pronoun of the form *isn't it?*, *can we?* or *would you?*, etc., is a much studied phenomenon in the field of linguistics (McGregor, 1995; Kimps, 2007). According to the Longman Grammar (Biber *et al.*, 1999), approximately 20% of the questions found in the conversational part of the corpus they analysed were tag questions.

Question tags are peripheral interrogative elements, typically appended to a declarative sentence, with the effect of modifying the sentence, to express a range of different attitudes, including doubt, surprise, contempt, etc. or simply to facilitate the flow of conversation. There are two main types of tag question, grammatical tag questions, also known as canonical tag questions, which are grammatically bound to the main anchor clause, for example:

(1) *You do believe in happy endings*, **don't you?**

(2) *He can't do that*, **can he?**

However, there also exist lexical (or invariant) tag questions, which are not grammatical bound in the same way, and are in fact the most frequent form of tag question, also commonly appearing in languages other than English. For example:

(7) *He's a proper bad man*, **innit?**

(8) *There's got to be a cure*, **right?**

The function of tag questions is complex and multi-facetted. They have often been referred to as dialogue facilitators, as ways of "keeping the conversation going" and "inviting listeners to

communicate" (Soars & Soars, 2000). However they have also been analysed as being modalisers to the main proposition of the utterance, modifying it to express the speaker's attitude to the proposition (McGregor, 1995). This bias based on the speaker's belief in the truth of the initial proposition is also linked to the understanding and common knowledge between speaker and interlocutor. The choice of question tag can have an important impact on the meaning of an utterance, in particular concerning speaker certainty, politeness and conduciveness. For example, a "they're a bit strange, aren't they?" suggests that the speaker thinks them to be strange, but demands confirmation from the listener in an inviting way, whereas "they're a bit strange, are they?" suggests that the listener has stated they are a bit strange and the speaker is questioning this, potentially in a slightly aggressive manner.

Tag questions are notably difficult to translate automatically because there are no direct equivalents to the English system of tag questions in other languages, with respect to canonical (grammatical) tag questions of the form "isn't it?","won't you?".[6] The form of tag questions may be considered relatively easy to predict, once an utterance is known to be a tag question, because the grammatical structure of a tag question in most cases echoes the auxiliary verb of the main clause. However, there is also the option to use a lexical tag question (e.g. "right?", "see?"), which complicates this decision. The most difficult decision is to know when to produce a tag question when translating from a language that uses different strategies for the same communicative purpose. SMT systems, which only use a very notion of local, immediate context, often struggle to provide a grammatical and coherent tag question. NMT systems theoretically far better, but the task of generating a tag question where one does not appear in the source sentence is still a considerable challenge.

4.1 Predicting the use of English tag questions: presence and form

To our knowledge, ours is the first work on the MT of tag questions in English, despite the vast linguistic literature on the topic. For the language direction French to English, we aim to improve the MT output of sentences whose English reference translations correspond to English tag questions, without degrading translation performance overall. We formalise the problem as a post-edition, classification problem on top of a strong SMT baseline. Given a high-performing phrase-based system (Koehn *et al.*, 2007), trained on parliamentary data, conference talks and subtitles, we modify the sentences classified as containing a tag question by appending the tag question form predicted by the classifier.

Tag question identification We first automatically annotate a subset of the OpenSubtitle parallel corpus (Lison & Tiedemann, 2016) for the presence or absence of tag questions on the English side of the corpus and for the form of the question tag if the subtitle is indeed a tag question. We use a sequence of heuristic, lexical-based rules based on our knowledge of English syntax and frequent patterns. A manual evaluation on a 500-subtitle subset of our annotations shows that precision and recall are approximately 98% and 100% respectively for the identification of sentence-final grammatical tag questions whose anchor clause is in the same subtitle. We divide the corpus consecutively into four datasets (train, dev1, dev2 and test), whose sizes are shown in Table 3. The number of different tag questions are also provided (corresponding to approximately 1% of sentences in each dataset). The distribution between the different question tag forms is very unequal (with the form "right" making up 20% of all tag question forms. This makes both learning and evaluation difficult.

[6]Grammatical tag questions do appear in some other languages, but this is rare (Axelsson, 2011).

	#sents	#TQs all	grammatical	lexical
train	18.5M	162,124 (0.9%)	71,889 (0.4%)	90,235 (0.5%)
dev1	5M	49,908 (1%)	15,070 (0.3%)	34,838 (0.7%)
dev2	5M	48,825 (1%)	13,931 (0.3%)	34,894 (0.7%)
test	5M	48,676 (1%)	13,212 (0.3%)	35,646 (0.7%)

Table 3: Distribution of English TQs in the four datasets.

Experimental setup We perform classification in two separate steps, first by predicting whether or not a tag question should be used in the target translation and secondly by predicting the true form of the question tag (e.g. "isn't it", "don't you think") for those sentences predicted to be tag questions by the first classifier. Both classifiers are linear classifiers trained using Vowpal Wabbit (Langford *et al.*, 2009).

The first classifier, whose task is to predict the presence or absence of a tag question, is trained on features extracted only from source data from the train set. The second classifier uses information from the current and following source sentences, the current and following target sentence and the score from the first classifier. Only those sentences to which the first classifier assigns a score above a certain threshold are sent to the second classifier, all other sentences being considered as non-tag questions. The value of this threshold at training time (which sentences are selected for inclusion into the second classifier's training set) and at testing time (which sentences will be eligible for being assigned a tag question by the second classifier) constitute the two hyperparameters of our system. The first classifier is trained on the training set, and the second classifier is trained on the subpart of the dev1 set selected by the first hyperparameter, both hyperparameters being optimised on the dev2 set. Features include significant bag of word features (unigram, bigram and trigram) filtered using a G^2 test, the n-grams of the sentence, the presence of a lexical tag in French and the presence of an affirmative or negative reply in the following subtitle.

	Step 1: TQ identification	Step 2: TQ labelling precision (%)			BLEU scores	
	F-score	ALL	TQ	Non-TQ	TQ	ALL
Baseline	48.3	99.02	29.56	99.7	31.26	34.11
Our system	62.6	99.15	20.94	99.9	32.08	34.12
Topline	-	1	1	1	46.68	34.33

Table 4: Post-edition results on the test set (f-score, precision and BLEU). Scores marked "TQ" apply only to sentences which contain a tag question in the reference translation. Scores marked "ALL" apply to all sentences. The baseline score is the machine translated output and the topline is the machine translated output, post-edited with the gold question tags.

Classification Results Results are shown in Table 4 for both classification steps. For the second step, we provide a total classification precision on all 214 labels (including the label "none" when no tag question is present).[7] However we also provide separate labelling precision for gold tag questions

[7]The second classifier is trained on the predicted outputs of the first classifier and so it is necessary to include the "none" label in the second step to allow the system to go back on the decision of the first classifier.

and for gold non-tag questions to give more insight into the performance of the classifier.[8] Finally, translation performance is evaluated using the automatic evaluation metric BLEU (Papineni *et al.*, 2002). In these preliminary experiments, we see that the overall labelling precision increases with respect to the baseline, but this is chiefly due to an increased precision in predicting when a subtitle is not a tag question. Our system predicts far more "non-TQ" labels than the baseline system, resulting in a lower precision for gold TQs. First observations show that our system has higher precision on those subtitles predicted as tag questions than the baseline, but this is at the expense of a lower recall (as reflected in the precision on gold TQs). We see slight increases in BLEU score for the entire corpus and on gold TQs. The lack of correlation between TQ precision and TQ BLEU should be investigated, and is possibly due to the fact that wrong question tag labels are penalised more than an absent tag label. Further experiments and a detailed manual evaluation would have to confirm this. The main problem with evaluating such a subjective, stylistic aspect is that the automatic evaluation metrics used here do not take into account the fact that several correct tag questions could be possible for a given sentence.

4.2 Analysis and perspectives

In future work, we aim to build on these preliminary experiments to improve the classification process and consequently the translation of tag questions. Possible perspectives include combining statistical classification for the identification step with robust rules for the prediction of the grammatical tag forms, which are most often very predictable given a main sentence.

Although tag question prediction may appear a very minor aspect of translation, appearing in only 1% of sentences, correctly predicting the tag question form could greatly improve the fluidity, coherence and naturalness of translated dialogues. In this experiment, we identified tag questions in the English subtitles using carefully constructed rules, but a future step could be to automatically identify such phenomena, which are difficult to translate, using the knowledge gained through a linguistic analysis of our corpus and an in-depth comparison of reference and baseline translations.

5 Conclusion

Context, in its many different forms, is crucial when translating speech-like texts. We have shown, through three different experiments on three aspects of contextual information, how integrating context into the translation process can help translation performance, whether during decoding as with our first experiment on domain adaptation or as a post-edition step as with the last two experiments. However, these methods remain limited. Domain adaptation by data partitioning restricts the amount of data used, pronoun predicting using external resources is, at present, limited by the lacunas in automatic coreference resolution, and we are far from being able to model the many different aspects of style, which, like tag questions, can greatly impact translation quality. The arrival of neural learning methods, which provide a promising architecture for taking into account richer linguistic information (Sennrich & Haddow, 2016), pave the way for integrating rich, contextual information into MT systems, and importantly, could provide a unified approach for the integration of the different types of context.

[8] TQ precision is the percentage of reference tag questions that are correctly labelled with the exact form of the tag question. Non-TQ precision is the percentage of reference non-tag questions that are correctly labelled.

References

AXELROD A., HE X. & GAO J. (2011). Domain Adaptation via Pseudo In-Domain Data Selection. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP'11, p. 355–362, Edinburgh, UK.

AXELSSON K. (2011). A cross-linguistic study of grammatically-dependent question tags. *Studies in Language*, **35**(4), 793–851.

BAWDEN R. (2016). Cross-lingual Pronoun Prediction with Linguistically Informed Features. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 564–570, Berlin, Germany.

BAWDEN R., WISNIEWSKI G. & MAYNARD H. (2016). Investigating gender adaptation for speech translation. In *Proc. of the 23rd Conférence sur le Traitement Automatique des Langues Naturelles*, TALN'16, p. 490–497, Paris, France.

BIBER D., JOHNANSSON S., LEECH G. & QUIRK R. (1999). *Longman Grammar of Spoken and Written English*. Longman, London.

BOHNET B. & NIVRE J. (2012). A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In *Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'12, p. 1455–1465, Jeju Island, Korea.

COATES J. (1986). *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*. London: Longman.

DABRE R., PUZIKOV Y., CROMIERES F. & KUROHASHI S. (2016). The Kyoto University Cross-Lingual Pronoun Translation System. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 571–575, Berlin, Germany.

FINCH A., SUMITA E. & NAKAMURA S. (2009). Class-Dependent Modeling for Dialog Translation. *IEICE Transactions on Information and Systems*, **92**(12), 2469–2477.

FOSTER G. & KUHN R. (2007). Mixture-Model Adaptation for SMT. In *Proc. of the 2nd Workshop on Statistical Machine Translation*, WMT'07, p. 128–135, Prague, Czech Republic.

GUILLOU L. (2016). *Incorporating Pronoun Function into Statistical Machine Translation*. PhD thesis, School of Informatics. University of Edinburgh.

GUILLOU L., HARDMEIER C., NAKOV P., STYMNE S., TIEDEMANN J., VERSLEY Y., CETTOLO M., WEBBER B. & POPESCU-BELIS A. (2016). Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 525–542, Berlin, Germany.

HARDMEIER C. (2012). *Discourse in statistical machine translation. a survey and a case study*. PhD thesis, Uppsala University, Uppsala, Sweden.

HARDMEIER C., NAKOV P., STYMNE S., TIEDEMANN J., VERSLEY Y. & CETTOLO M. (2015). Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proc. of the 2nd Workshop on Discourse in Machine Translation*, DiscoMT'15, p. 1–16, Lisbon, Portugal.

HOVY D. (2015). Demographic Factors Improve Classification Performance. In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL-IJCNLP'15, p. 752–762, Beijing, China.

KIMPS D. (2007). Declarative constant polarity tag questions: A data-driven analysis of their form, meaning and attitudinal uses. *Journal of Pragmatics*, p. 270–291.

KOEHN P. & HOANG H. (2007). Factored translation models. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 868–876, Prague, Czech Republic.

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. & HERBST E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, ACL'07, p. 177–180, Prague, Czech Republic.

KUHN R., NGUYEN P., JUNQUA J.-C., GOLDWASSER L., NIEDZIELSKI N., FINCKE S. & CONTOLINI M. (1998). Eigenvoices for speaker adaptation. In *Proc. of the 5th International Conference on Spoken Language Processing*, IWSLT'98, p. 1771–1774, Sydney, Australia.

LAKOFF R. (1975). *Language and woman's place: text and commentaries*. New York: Oxford University Press.

LANGFORD J., LI L. & ZHANG T. (2009). Sparse Online Learning via Truncated Gradient. *The Journal of Machine Learning Research*, p. 777–801.

LAVIE A. & DENKOWSKI M. J. (2009). The Meteor metric for automatic evaluation of machine translation. *Machine Translation*, **23**(2-3), 105–115.

LE NAGARD R. & KOEHN P. (2010). Aiding pronoun translation with co-reference resolution. In *Proc. of the 5th Workshop on Statistical Machine Translation*, WMT'10, p. 252–261, Uppsala, Sweden.

LISON P. & TIEDEMANN J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proc. of the 10th Language Resources and Evaluation Conference*, LREC'16, p. 923–929, Portorož, Slovenia.

LUOTOLAHTI J., KANERVA J. & GINTER F. (2016). Cross-lingual pronoun prediction with deep recurrent neural networks. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 596–601, Berlin, Germany.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL'14, p. 55–60, Baltimore, USA.

MCGREGOR W. (1995). The English 'tag question': A new analysis, is(n't) it? *On Subject and theme: A discourse functional perspective*, **118**, 91–121.

MIRKIN S., NOWSON S., BRUN C. & PEREZ J. (2015). Motivating Personality-aware Machine Translation. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP'15, p. 1102–1108, Beijing, China.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, ACL'02, p. 311–318, Philadelphia, USA.

PECINA P., TORAL A. & VAN GENABITH J. (2012). Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation. In *Proc. of the 24th International Conference on Computational Linguistics*, COLING'12, p. 2209–2224, Mumbai, India.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

ROY A., GUINAUDEAU C., BREDIN H. & BARRAS C. (2014). TVD: A Reproducible and Multiply Aligned TV Series Dataset. In *Proc. of the 9th Language Resources and Evaluation Conference*, LREC'14, p. 418–425, Reykjavik, Iceland.

SAGOT B. (2010). The Le*fff*, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proc. of the 7th International Conference on Language Resources and Evaluation*, LREC'10, p. 2744–2751, Valletta, Malta.

SALUJA A., LANE I. & ZHANG Y. (2011). Context-aware Language Modeling for Conversational Speech Translation. In *Proc. of the 13th Machine Translation Summit*, p. 97–104, Xiamen, China.

SENNRICH R. & HADDOW B. (2016). Linguistic Input Features Improve Neural Machine Translation. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 83–91, Berlin, Germany.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'16, p. 35–40, California, USA.

SOARS J. & SOARS L. (2000). *New headway English course. Pre-intermediate student's book*. Oxford: Oxford University Press, 2nd edition.

STYMNE S. (2016). Feature exploration for cross-lingual pronoun prediction. In *Proc. of the 1st Conference on Machine Translation*, WMT'16, p. 609–615, Berlin, Germany.

TIEDEMANN J. (2016). A Linear Baseline Classifier for Cross-Lingual Pronoun Prediction. In *Proc. of the 1st Conference on Machine Translation*, p. 616–619, Berlin, Germany.

VAN DER WEES M., BISAZZA A. & MONZ C. (2016). Measuring the Effect of Conversational Aspects on Machine Translation Quality. In *Proc. of the 26th International Conference on Computational Linguistics*, COLING'16, p. 2571–2581, Osaka, Japan.

VOLKOVA S., WILSON T. & YAROWSKY D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP'13, p. 752–762, Beijing, China.

WANG L., ZHANG X., TUY Z., WAY A. & LIU Q. (2016). Automatic construction of discourse corpora for dialogue translation. In *Proc. of the 10th International Conference on Language Resources and Evaluation*, LREC'16, p. 2748–2754, Portorož, Slovenia.