Machine Translation, it's a question of style, innit? The case of English tag questions

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

Abstract

In this paper, we address the problem of generating English tag questions (TQs) (e.g. *it is,* ***isn't it?***) in Machine Translation (MT). We propose a post-edition solution, formulating the problem as a multi-class classification task. We present (i) the automatic annotation of English TQs in a parallel corpus of subtitles and (ii) an approach using a series of classifiers to predict TQ forms, which we use to post-edit state-of-the-art MT outputs. Our method provides significant improvements in English TQ translation when translating from Czech, French and German, in turn improving the fluidity, naturalness, grammatical correctness and pragmatic coherence of MT output.

1 Introduction

When it comes to the machine translation (MT) of discourse, revisiting the question of what constitutes a high quality translation is essential; which aspects of language should be tackled and how to evaluate them. A first step is to identify the many stylistic aspects of speech that pose a problem for current MT techniques and to study how they could be taken into account and evaluated.

We take a step in this direction by addressing a new aspect of discourse in MT, related to speaker attitude and style: the English tag question (hereafter TQ), i.e. utterances such as *catchy,* ***ain't it?*** and *it wasn't him,* ***was it?***. When translating into English, TQs present two main challenges. The first is knowing when to generate one. Similar to the translation of discourse connectives, TQ use is a question of style and speaker attitude, and importantly, there is not often a direct, lexical correspondence across languages. TQs are common in English and far less so in other languages, which means that other contextual cues are necessary to determine whether a TQ should appear in the English translation. The second, in particular for canonical TQs (*e.g. was it?, isn't it?*), is that the overall grammaticality of the utterance is determined by the correct choice of tag, which, in a similar way to anaphor translation, is grammatically dependent on the rest of the MT output.

Our aim in this paper is to improve the generation of English TQs in an MT setting with English as the target language. We formulate this as a multi-class classification task, using features from both source sentences and machine translated outputs. The prediction of the appropriate question tag to use in the English translation (if any) is then used to post-edit MT outputs. Our results, when translating from Czech (CS), French (FR) and German (DE) into English (EN), display significant improvements, as shown by automatic and manual evaluations (Sec. 5).[1]

2 English tag questions (TQs)

TQs are interrogative constructions, common in spoken English, formed of a main clause (typically declarative), followed by a peripheral interrogative element, the *question tag*:

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

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

In its canonical form, the English question tag (in bold) is formed of an auxiliary verb, which can be negated, followed by a pronoun. It parallels the verb and subject (underlined) of the preceding host clause (in italics). The grammatical structure of TQs and agreement between the host and the tag gives them the name of *grammatical TQs*.[2]

[1] All scripts and annotations are freely available at http://diamt.limsi.fr.

[2] Although in theory their form is relatively systematic, their attested usage is more complex.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2497–2502
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

There exists a second type of TQ, *lexical TQs*, formed of a word or phrase that is invariant to the subject and verb of the host clause. For example:

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

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

TQs' functions are complex and communicate information about speaker attitude, tone, the relationship between dialogue participants, common ground and dialogue flow (McGregor, 1995). Yet few languages have such a systematic use of TQs, in particular of grammatical TQs, as English (Axelsson, 2011). When translating into English, the first difficulty is appropriately generating English TQs from a source sentence that does not have a TQ; the complex and often ambiguous functions of TQs (e.g. expressing doubt or surprise) can be expressed differently, and subtly, in the source language. The second difficulty is ensuring grammatical coherence of canonical TQs. Consider the German sentence *Sie lebt noch, nicht wahr?* and its English translation *She's alive, isn't she?*. The choice of the question tag *isn't she?* is dependent on the subject and verb of the anchor clause. Had the translation been *She still lives*, the correct question tag would have been *doesn't she?*.

3 Related work

Discourse is a growing field in MT (Le Nagard and Koehn, 2010; Hardmeier, 2012, 2014). To our knowledge, there has been no previous work on TQs in MT, but the two main challenges described above are similar to those associated with two previously studied discursive aspects: the translation of discourse connectives and of anaphoric pronouns. As with TQs, their frequency is relatively low, but their mistranslation has a high impact on coherence, naturalness and therefore human understanding of translations (Meyer and Popescu-Belis, 2012).

Discourse connectives (e.g. *since*, *because*) indicate the relation between discursive units and are linked to the overall coherence of a text in a similar way to TQs. They often have no direct mapping when translated (Meyer and Webber, 2013) and it is often necessary to generate a discourse connective or a TQ where one is not present in the source sentence. Previous work by Meyer and Webber (2013) consists in the disambiguation of discourse connectives in source sentences prior to translation, using automatic sense classification, which guides the MT system's choice of how a discourse connective should be translated (if at all). However they do not handle the case of generating discourse relations from a source sentences in which they do not appear lexically, as is our aim for TQs.

The difficulty of anaphoric pronoun translation is ensuring grammatical agreement between a pronoun and its coreferent, when the information relevant to grammatical agreement in the target language is not present in the source language. For example, the French translation of the pronoun *it* in *I hear an owl but I can't see it* is translated as *le* or *la*, depending on whether *owl* is translated as masc. *hibou* or fem. *chouette*. The position of the coreferent, as with the subject and verb for TQs, is not pre-determined, and identifying which words the translated pronoun or TQ must agree with is not always easy. The majority of works perform classification of pronominal forms in view to post-editing MT output (Guillou, 2016), encouraged by the shared task on cross-lingual pronoun prediction at DiscoMT15 (Hardmeier et al., 2015) and WMT16 (Guillou et al., 2016). We use the same strategy here, since the coherent use of TQs, like anaphors, is dependent on the MT translation.

4 Our post-edition approach

Our method to improve the generation of English TQs in MT is the automatic post-edition of state-of-the-art MT outputs. We formulate the problem as a supervised, multi-class classification task, exploiting lexical features from source sentences and their machine translations (Sec. 4.2). Possible labels are the different question tag forms (e.g. *isn't it*, *ok*). Predicted tags are either used to replace the question tag already present in the MT output or appended to it otherwise. We test our method for three source languages (CS, DE and FR) into EN.

4.1 Corpus annotation for English TQs

The first step is to produce annotated data. The corpora used cover three language pairs: CS-EN, DE-EN and FR-EN, and are large subsets of the most recent films of the OpenSubtitles[3] parallel corpus (Lison and Tiedemann, 2016). The subtitles were automatically cleaned using heuristics and processed with the MElt tokeniser (Denis and Sagot, 2012) and the Moses truecaser (Koehn et al., 2007). We then developed robust, manually defined lexical rules to identify English TQs. We

[3] www.opensubtitles.org

identify both the presence of TQs and the question tags themselves (e.g. *is it?*, *right?*).

A manual evaluation on a random subset of 500 grammatical TQs and 500 lexical TQs shows that annotations are near perfect (accuracy of ≈98% and recall of 100% on sentence-final grammatical TQs whose host clause is in the same subtitle).[4]

Each corpus was divided into three sets: TRAIN ($\frac{2}{3}$), DEV ($\frac{1}{6}$) and TEST ($\frac{1}{6}$). The distribution of TQs is shown in Tab. 1. TQs make up approximately 1% of subtitles, but are common among questions (≈20%). There are between 238 and 285 distinct English question tags, depending on the language pair (including a label *none* for non-TQs). The most frequent question tag is *right?* (≈20%), followed by *ok?* (≈16%). The majority of labels are grammatical question tags, but the most frequent (*isn't it?*) represents only ≈3% of all TQs, revealing a huge class imbalance.

	#sents	#English TQs all	gram.	lex.	#labels gram.	lex.
CS-EN	15.1M	146,782	44,572	102,210	269	15
DE-EN	6.2M	57,435	18,396	39,039	221	16
FR-EN	15.1M	149,847	44,651	105,196	254	16

Table 1: TQ distribution for each language pair.

4.2 Question tag classification

Given the class imbalance (Sec. 4.1) and the different nature of lexical and grammatical TQs, we hypothesise that first predicting the presence of an English TQ before selecting tag forms is preferable to directly predicting tags in a single, direct pass. We compare a single statistical classifier (hereafter CL-ONE), which directly predicts the question tag (including the label *none*), to a more complex system using a sequence of classifiers (hereafter CL-SEQ, see Fig. 1). In CL-SEQ, a first classifier (CL-SEQ$_{\text{coarse}}$) predicts a coarse-grained label *gram* (grammatical TQ), *lex* (lexical TQ) or *none* (non-TQ), which determines which classifier (CL-SEQ$_{\text{lex}}$ or CL-SEQ$_{\text{gram}}$) is used to predict the tag form. Both CL-SEQ$_{\text{coarse}}$ and CL-SEQ$_{\text{lex}}$ are statistical classifiers. CL-SEQ$_{\text{lex}}$ is trained on the TRAIN examples assigned the label *lex* by CL-SEQ$_{\text{coarse}}$, which explains why it provides a second chance to predict grammatical or non-TQs. CL-SEQ$_{\text{gram}}$ is a rule-based system, a choice that is better adapted to the sparse labels of grammatical TQs. Where there is no rule available, this system too can predict the label *none*.

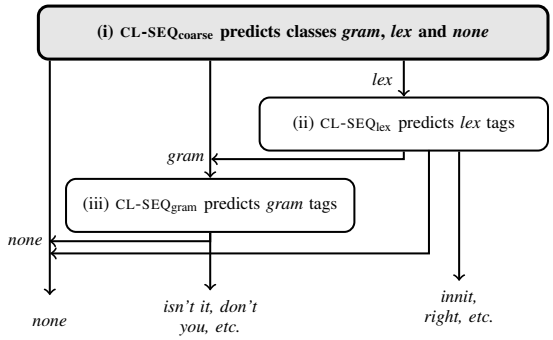

Figure 1: CL-SEQ classification: (i) into coarse-grained classes, (ii–iii) prediction of forms

Experimental setup All statistical classifiers are linear classifiers trained using Vowpal Wabbit (Langford et al., 2009).[5,6] To account for class imbalance, examples are weighted according to their relative frequency in the TRAIN set, and the degree of weighting is optimised on the DEV set.[7] Features used for all statistical classifiers are described just below. The rule-based CL-SEQ$_{\text{gram}}$ system relies on the MT output alone.

Features We use automatically and manually defined lexical feature templates. Unless indicated the features apply to both the source sentence and the MT output. The first set of features are automatically identified bag-of-word features, which represent the 500 uni-, bi- and tri-grams most associated with a TQ, as measured by a G^2 test. The second set of features are manually defined, based on language-specific question-response patterns and recognisable lexical clues. They include (i) the presence of a question tag (and its form), (ii) the presence of a final question mark, (iii) (CS and DE only) whether the following subtitle contains a verb that appears in the current subtitle (and if so, we include as a feature the verb type and the preceding word in both the current and following subtitles)[8], (iv) the following subtitle contains a specific response (from a predefined list of replies

[4] Recall for lexical TQs cannot be accurately measured, as identification relies on a closed list of forms found in the literature and observed in the data.

[5] http://hunch.net/~vw/

[6] We use "OAA", FTRL-proximal optimisation, L2 regularisation ($\lambda = 1e-6$) and quadratic features.

[7] We vary weights from equal for all examples to weights that fully counterbalance the class distribution.

[8] In German and Czech, it is common for a reply to a yes/no question to repeat the verb of the question, e.g. *Poslala jsi mu to?* 'Did you send it to him?' — *Poslala jsem* 'Yes, I did' (lit. 'send (I)_did') (Gruet-Skrabalova, 2013).

Lang. pair		Gram TQs				Lex TQs				Non-TQs			Overall
		P	R	F	P*	P	R	F	P*	P	R	F	P*
CS→EN	baseline	52.22	43.83	47.66	35.75	49.76	57.51	53.35	45.35	99.69	99.63	99.66	99.11
	CL-ONE	66.15	15.57	25.20	50.09	55.19	40.33	46.60	51.20	99.43	99.84	99.63	99.16
	CL-SEQ	56.87	44.96	50.22	38.85	60.76	46.68	52.79	56.58	99.57	99.79	99.68	99.21
DE→EN	baseline	45.72	28.68	35.25	9.97	69.07	45.16	54.61	66.59	99.51	99.83	99.67	99.21
	CL-ONE	69.76	9.42	16.59	48.54	61.63	45.49	52.34	59.78	99.46	99.88	99.67	99.26
	CL-SEQ	59.27	42.74	49.67	35.48	68.70	53.21	59.97	66.69	99.62	99.84	99.73	99.32
FR→EN	baseline	41.15	47.18	43.96	12.95	57.63	38.25	46.18	52.55	99.53	99.72	99.62	99.03
	CL-ONE	66.30	9.36	16.41	44.87	55.05	28.80	37.81	51.73	99.32	99.89	99.60	99.12
	CL-SEQ	58.48	33.95	42.96	38.22	63.02	38.22	47.59	59.42	99.46	99.85	99.65	99.19

Table 2: Precision (P), Recall (R), F-score (F) and fine-grained labelling precision (P*) for the TEST set on each language pair. Results are given for each coarse-grained TQ class (*gram*, *lex* and *non-TQ*). Labelling precision is calculated on the subtitles with the corresponding predicted coarse-grained label. Marked in grey are the cells containing the best F-scores for coarse-grained label groupings and the overall labelling precision (for fine-grained classes).

such as *OK*, *yes*, *no*, etc.), and (v) the first words of the MT output (1-4 gram), the last auxiliary, the last pronoun and the last pronoun-auxiliary pair.

Rule-based grammatical TQ prediction Our rule-based approach is designed to predict which grammatical tag should be appended to a given translation. The rules consist in the identification of certain lexical cues from the translation. For instance, utterance-initial words can be a good indicator of the use of a particular question tag: imperatives such as *let's ...* (indicative of the tag *shall we*), and claims about the interlocutor's perception such as *you think...* or *you know...* (indicative of the tag *don't you*). When there is a single auxiliary and subject, these are directly used to construct a question tag, using, as a simplification, the opposite polarity to that of the anchor clause, which is the most common polarity pattern in TQs. We include several rules to account for complex clauses and perform grammatical checking between the subject and auxiliary of the question tag. The complete set of rules is available at the address cited in Footnote 1.

"Baseline" MT outputs For Czech and German, we use the top systems at WMT16, both attentional encoder-decoder NMT models (Sennrich et al., 2016). For French, we trained a phrase-based model with Moses (Koehn et al., 2007).[9] Baseline predictions are automatically extracted from the MT outputs using our English TQ identification rules (Sec. 4.1).

[9]We use a combination of three phrase tables and three 4-gram KenLM language models (Heafield et al., 2013), trained on Europarl, Ted Talks and 3M-sentence subtitles, tuned using `kbmira` on a disjoint 2.5K-sentence subset.

5 Results and analysis

As mentioned by Hardmeier (2012), evaluating coherence-related MT phenomena is problematic. A question tag can be the correct choice without matching the question tag form in the reference translation (Sec. 2), making traditional metrics involving lexical comparison (including all standard MT evaluation metrics) ill-adapted to the task.

Despite this, we provide in Tab. 2 results using traditional metrics. To get a better view of the scores, we group final predicted question tags into their coarse-grained classes (*gram*, *lex*, *none*) and calculate the precision (P), recall (R) and F-score (F) for these three classes. Within each coarse-grained class, we also provide labelling precision (P*), corresponding to the number of question tags within that coarse-grained class assigned the correct question tag form according to the gold label. Labelling precision is also given overall for all test sentences (in the final column).

Overall labelling precision is significantly improved for all language pairs with both classification systems, but in particular for CL-SEQ. This is partly due to a better prediction of non-TQs, represented by the high corresponding F-scores for CL-SEQ for all three language pairs. However it is also linked to a better labelling of grammatical and lexical TQs, which can be seen by the high labelling precision (P*) in the context of high recall (R). The higher scores of CL-SEQ over CL-ONE, particularly in terms of recall, which are most likely a result of question tag label sparsity, show that our two-tier strategy of predicting grammatical and lexical tags separately is better adapted than a single classifier.

There is a notable drop in recall between CL-

Source	Reference	Baseline	CL-SEQ	Judgement
Du hast das gemacht, nicht wahr?	You did this , **didn't you?**	You've done that, **don't you?**	haven't you?	Improved
Das Gehirn zerstören, wisst ihr?	Kill the brain, **you know?**	Destroying the brain, **you know?**	none	Degraded
Das stimmt, oder?	I know, **right?**	That's true, **right?**	isn't it	Equal (correct)
Das ist merkwürdig, oder?	It's weird, **isn't it?**	That 's odd, **does it?**	none	Equal (incorrect)

Table 3: Some examples from the manual comparison of the baseline translation and CL-SEQ predictions. An example is given for each of the possibilities: the prediction is better, worse or the translation and the prediction are equally good or poor with respect to tag question prediction.

SEQ and CL-SEQ when it comes to grammatical TQ prediction, which is not as marked for lexical TQ prediction. This is most likely due to the huge class imbalance in the different question tags (221 grammatical tags vs. 16 lexical tags), which causes the purely statistical one-pass system to favour the more frequent lexical tags and struggle to predict the wide range of much rarer grammatical tags.

		baseline			
		gram	lex	none	Total
gold	gram	871	180	1986	3037
	lex	429	2878	3066	6373
	none	605	1109	1019408	1021122

		predicted			
		gram	lex	none	Total
gold	gram	1298	367	1372	3037
	lex	418	3391	2564	6373
	none	474	1178	1019470	1021122

Table 4: Confusion matrix of for baseline and predicted versus gold tags (when question tags are grouped into their three coarse-grained classes) for DE→EN.

Tab. 4 shows a comparison of baseline and predicted tags for the DE→EN test set. For ease of illustration, the question tags are again regrouped into their three coarse-grained classes (*gram*, *lex* and *none*). The matrix reveals that for predicted lexical tags and non-TQs, the majority were correctly classed into these coarse-grained classes. However, grammatical tags proved more difficult to predict, the majority being classed as non-TQs, most likely a result of the fact that no such tag question was present on the German source side. The most common errors on this test set were predicting a non-TQ when a lexical tag was expected. For example, CL-SEQ predicted *none* 1112 times when the gold tag *right?* was expected. However the number of correct predictions of *right* exceeded this number (1371 cases) and compared to baseline predictions, all coarse-grained categories show an improvement in recall.

Manual analysis Given the drawbacks of automatic metrics, we manually evaluated a set of final translations post-edited with the predictions produced by the CL-SEQ system for DE→EN. We randomly selected 100 examples for which the baseline translation was modified and, comparing the baseline and CL-SEQ prediction, we labelled the example as *improved* (baseline incorrect and prediction correct) or *degraded* (baseline correct and prediction incorrect) or *equal* (both baseline and prediction (in)correct). Examples of these choices are provided in Tab. 3. We found that post-edition made an improved choice of tag in 59/100 examples (compared to 13 worse choices). Only 25 of the 59 improved examples exactly matched the reference tag, confirming the problem of relying solely on traditional metrics. We notice in particular an improvement in the choice of grammatical TQs (33 out of the 59 improvements).

6 Discussion and perspectives

Improvements in the generation of English TQs in MT outputs, as seen by our manual analysis, result in improved grammatical coherence, particularly for grammatical TQs. However, TQ translation is far from solved. As a stylistic aspect, its prediction and evaluation are complex and should be further explored. Possible improvements include improving the choice of linguistic information used and using this work to explore how TQs' functions are portrayed in languages other than English.

Interesting future work would be to compare the opposite approach to the task; augmenting source sentences with disambiguating information prior to translation, particularly within an NMT framework, which has good potential for handling non-local context and integrating extra features.

References

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

Pascal Denis and Benoît Sagot. 2012. Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, 46(4):721–736.

Hana Gruet-Skrabalova. 2013. Verbs and particles in minimal answers to yes-no questions in Czech. In *Formal Description of Slavic Languages 10*, Slavic Grammar from a Formal Perspective, the 10th anniversary FDSL conference, pages 197–215, Leipzig, Germany.

Liane Guillou. 2016. *Incorporating Pronoun Function into Statistical Machine Translation*. Ph.D. thesis, School of Informatics. University of Edinburgh.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the 1st Conference on Machine Translation*, WMT'16, pages 525–542, Berlin, Germany.

Christian Hardmeier. 2012. Discourse in statistical machine translation. a survey and a case study. *Discours [online]*, 11.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. Ph.D. thesis, Uppsala University, Department of Linguistics and Philology, Uppsala, Sweden.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation*, DiscoMT'15, pages 1–16, Lisbon, Portugal.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 690–696, Sofia, Bulgaria.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, ACL'07, pages 177–180, Prague, Czech Republic.

John Langford, Lihong Li, and Tong Zhang. 2009. Sparse Online Learning via Truncated Gradient. *The Journal of Machine Learning Research*, pages 777–801.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the 5th Workshop on Statistical Machine Translation*, WMT'10, pages 252–261, Uppsala, Sweden.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th Language Resources and Evaluation Conference*, LREC'16, pages 923–929, Portorož, Slovenia.

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

Thomas Meyer and Andrei Popescu-Belis. 2012. Machine Translation of Labeled Discourse Connectives. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas*, pages 129–138, San Diego, California, USA.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the 1st Workshop on Discourse in Machine Translation*, DISCOMT '13, pages 19–26, Sofia, Bulgaria.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the 1st Conference on Machine Translation*, WMT '16, pages 368–373, Berlin, Germany.