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

To cite this version:

HAL Id: hal-01588171
https://hal.science/hal-01588171
Submitted on 15 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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).

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*.1

---

1Although in theory their form is relatively systematic, their attested usage is more complex.
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 coherence 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-editing 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\(^2\) 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

\(^2\)www.opensubtitles.org
defined lexical rules to identify English TQs. We 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).

Each corpus was divided into three sets: TRAIN (2/3), DEV (1/6) and TEST (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.

<table>
<thead>
<tr>
<th></th>
<th>#sents</th>
<th>all</th>
<th>#English TQs</th>
<th>#labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>gram. lex. gram. lex.</td>
<td></td>
</tr>
<tr>
<td>CS-EN</td>
<td>15.1M</td>
<td>146,782</td>
<td>44,572</td>
<td>102,210</td>
</tr>
<tr>
<td>DE-EN</td>
<td>6.2M</td>
<td>57,435</td>
<td>18,396</td>
<td>39,039</td>
</tr>
<tr>
<td>FR-EN</td>
<td>15.1M</td>
<td>149,847</td>
<td>44,651</td>
<td>105,196</td>
</tr>
</tbody>
</table>

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\textsubscript{coarse}) predicts a coarse-grained label *gram* (grammatical TQ), *lex* (lexical TQ) or *none* (non-TQ), which determines which classifier (CL-SEQ\textsubscript{lex} or CL-SEQ\textsubscript{gram}) is used to predict the tag form. Both CL-SEQ\textsubscript{coarse} and CL-SEQ\textsubscript{lex} are statistical classifiers. CL-SEQ\textsubscript{lex} is trained on the TRAIN examples assigned the label *lex* by CL-SEQ\textsubscript{coarse}, which explains why it provides a second chance to predict grammatical or non-TQs. CL-SEQ\textsubscript{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*.

**Experimental setup** All statistical classifiers are linear classifiers trained using Vowpal Wabbit (Langford et al., 2009). 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. Features used for all statistical classifiers are described just below. The rule-based CL-SEQ\textsubscript{gram} system relies on the MT output alone.

**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.

---

3All scripts and annotations will be made freely available in order to be able to recreate the corpus.

4Recall 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.

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

6We use “OAA”, FTRL-proximal optimisation, L2 regularisation (λ = 1e−6) and quadratic features.

7We vary weights from equal for all examples to weights that fully counterbalance the class distribution.
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 recognizable 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), (iv) the following subtitle contains a specific response (from a predefined list of replies 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.

"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). Baseline predictions are automatically extracted from the MT outputs using our English TQ identification rules (Sec. 4.1).

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 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).

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).

<table>
<thead>
<tr>
<th>Lang. pair</th>
<th>Gram TQs</th>
<th>Lex TQs</th>
<th>Non-TQs</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
<td>P*</td>
</tr>
<tr>
<td>CS→EN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>92.22</td>
<td>43.83</td>
<td>47.66</td>
<td>35.75</td>
</tr>
<tr>
<td>CL-ONE</td>
<td>66.15</td>
<td>15.57</td>
<td>25.20</td>
<td>50.09</td>
</tr>
<tr>
<td>CL-SEQ</td>
<td>56.87</td>
<td>44.96</td>
<td>50.22</td>
<td>38.85</td>
</tr>
<tr>
<td>DE→EN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>45.72</td>
<td>28.88</td>
<td>35.25</td>
<td>9.97</td>
</tr>
<tr>
<td>CL-ONE</td>
<td>69.76</td>
<td>9.42</td>
<td>16.59</td>
<td>48.54</td>
</tr>
<tr>
<td>CL-SEQ</td>
<td>59.27</td>
<td>42.74</td>
<td>49.67</td>
<td>35.48</td>
</tr>
<tr>
<td>FR→EN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>41.15</td>
<td>47.18</td>
<td>43.96</td>
<td>12.95</td>
</tr>
<tr>
<td>CL-ONE</td>
<td>66.30</td>
<td>9.36</td>
<td>16.41</td>
<td>44.87</td>
</tr>
<tr>
<td>CL-SEQ</td>
<td>58.48</td>
<td>33.95</td>
<td>42.96</td>
<td>38.22</td>
</tr>
</tbody>
</table>
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 rarer grammatical tags.

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

<table>
<thead>
<tr>
<th></th>
<th>predicted</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>gram</td>
<td>lex</td>
<td>none</td>
</tr>
<tr>
<td>gold</td>
<td>1242</td>
<td>351</td>
<td>1341</td>
</tr>
<tr>
<td>gram</td>
<td>387</td>
<td>3370</td>
<td>2719</td>
</tr>
<tr>
<td>lex</td>
<td>2050</td>
<td>4896</td>
<td>1023586</td>
</tr>
<tr>
<td>none</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tab. 4 shows predicted tags against gold 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 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 1166 times when the gold tag right? was expected. However the number of correct predictions of right exceeded this number (1352 cases).

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-editing 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.

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


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.


