International Conference; Meaning in Translation: Illusion of Precision, MTIP2016, 11-13 May 2016, Riga, Latvia

Yes-No Answers, Partial Pro-drop Languages and Machine Translation

Eduardo Correa Soares*

Université Paris Diderot and CAPES Foundation, 8e étage, Batiment Olympe de Gouges, 8 rue Albert Einstein, Paris, 75013, France

Abstract

This paper addresses the automatic translations of verbal answers to yes-no questions from partial pro-drop languages (Brazilian Portuguese and Russian) into a non-pro-drop language (English). The outputs provided by standard statistical machine translations are mostly grammatically inaccurate or semantic-pragmatically inadequate. This paper proposes a question under discussion based annotation to improve the statistical correspondence. The results show the accuracy of the outputs was significantly increased as regards fidelity, adequacy and grammaticality.

Keywords: Verbal answers; statistical machine translation; discourse structure; partial pro-drop languages.

Introduction

This paper addresses verbal answers to yes-no questions[vYNAs] in automatic translations from two partial pro-drop languages (Brazilian Portuguese[BP] and Russian[RU]) into a non-pro-drop language (English[EN]) (Rizzi 1981, Chomsky 1981, 1986, Biberauer et al. (2010), Holmberg 2015, inter alia). In BP, if, for example, one asks permission to go to the toilet – as in (1.A) –, a proper answer can be a verb in the 2nd person singular, Pode “may” –such as in (1.B).

(1) A. – Posso ir ao banheiro? B. – Pode, ué.
    may-p-1s go-inf to.the toilet      may-p-2s, [Discursive Marker]

* Corresponding author. Tel.: +33157277801.
E-mail address: soares_ec@yahoo.com.br
“May I go to the toilet?”

“Yes, you may.”

(OPUS Corpora – Central do Brasil)

The same holds for RU: a possible answer to a negative question in RU – such as in (2A) is a negation (He “not”) and 3rd person singular verb – (2B). 1

(2) A. – He ест?

not eat-p-3s

“Is he/she not eating?”

B. – He ест.

not eat-p-3s

“No, he/she is not.”

(OPUS Corpora – Дневной дозор)

Such vYNAs in BP and RU are a hurdle for statistical machine translation (SMT), as observed in Fig. 1-2 because (i) null subjects are not translated into overt pronouns in the target language and (ii) the semantic-pragmatic correspondence between answers are not quite straightforward. The outputs in EN provided by the baseline SMT (for instance, Google®) are thus not semantic-pragmatically correspondent and are incidentally ungrammatical.

In the literature about machine translation, it is claimed that evaluations must take into account many aspects of translation, including fidelity, adequacy, grammaticality and fluency of the output (White & O’Connell 1994, Hovy 1999, inter alia). Since at least Sadock & Zwick (1985), it is also well-known in theoretical linguistics that the range of variation concerning answers to yes-no questions is large (see also Holmberg 2015, and the references cited there). In this paper, I demonstrate in detail how translations of vYNAs in BP and RU into EN are a shortcoming for SMTs. I suggest incorporating a pragmatic-based tagging system to fine-tune the statistical correspondence and further improve upon the output translations.

This paper is organized as follows. In section 2, I consider the difficulties that translations vYNAs from BP and RU into EN impose to SMTs. In section 3, I present and discuss some relevant literature. In section 4, I put forth my proposal and the ideas upon which I have built it. In section 5, I introduce the methodology used here and, in the following section, I show the results obtained. In the last section I discuss the results and forthcoming research.

2. Problem

Purely SMTs are based on frequency and bilingual texts (see Dorr et al., 1999 for an overview). Pairs of texts in two different languages are compared and statistically relevant correspondence is identified by an algorithm (Och, & Nei, 2003). For instance, the most popular MT, Google Translate®, does not apply any rules or rule-based grammatical analyses, since its creator has criticized the accuracy and usefulness of rule-based algorithms (Och 2005). There is the rub: verbs as those of vYNAs are statistically relevant in many other linguistic environments, where they are simply translated into an inflectioned verb. In Table 1, I show the distribution of one lexical item (poder “can/could/shall”) in the source language (BP) as regards pairing with lexical items in the target language (EN) in one movie subtitle. In 32 occurrences of this item, 23 are paired with a verb. In column class, I present the distribution of this item according to the classification of paired item in the target language: auxiliary of future, imperative mode, modal verb, other and vYNA. As shown in Table 2, the verb poder is used in several contexts, being significantly more translated into modal verbs than into vYNAs in EN. This explains why the translation provided by the baseline SMT (Fig. 1) is the bare verb can in EN. Being paired this way, the output is ungrammatical.

![Fig. 1. A translation of a positive vYNA from BP into EN.](image1)

![Fig. 2. A translation of a negative vYNA from RU into EN.](image2)

1 In (2), I have translated the verbs in the present tense in RU into the progressive in EN for contextual reasons. As shown in the Fig. 2 below example (2), the baseline SMT does not do so. Such a divergence is out of the scope of this paper. Hereafter I ignore problems with tense.
Table 1. Word-based pairing of lexical items.

<table>
<thead>
<tr>
<th>Classification of the paired item in the output</th>
<th>Number of occurrences</th>
<th>Relative weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>modal_verb</td>
<td>14</td>
<td>0.4375</td>
</tr>
<tr>
<td>imperative_mode</td>
<td>6</td>
<td>0.1875</td>
</tr>
<tr>
<td>other</td>
<td>5</td>
<td>0.1562</td>
</tr>
<tr>
<td>vYNAs</td>
<td>4</td>
<td>0.125</td>
</tr>
<tr>
<td>auxiliar_future</td>
<td>3</td>
<td>0.0937</td>
</tr>
<tr>
<td>total</td>
<td>32</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. The verb poder in BP translated into EN.

In a larger corpus, the algorithm will generate a correlation of this verb with a verbal output, since the verb pode will be a number of times paired with the modal correspondents in EN. When this word is introduced as an input for a SMT, the algorithm will call the most statistically relevant correspondence, giving a modal verb as the output. Beyond generating ungrammatical sentences, there is an additional shortcoming of SMTs concerning the proper treatment of translations of vYNAs from BP and RU (and many other languages with similar system of answers) into EN (and others with comparable systems). As pointed out by Valduví (1990), Lambrecht (1994), among many others, many sentences in a given language can be uttered in several different forms as regards word-order, elided constituents, etc. even though they have the same core meaning. The reasons for such a variation are information (or discourse) structure demands. These “allosentences” do not correspond to the same sentence into different languages, because a same information structure package can be encoded into different grammatical constructions depending on the language.² Note, for instance, the relevant case in (3) below.

² Roughly allosentences are sentences whose proposition is the same, but whose shapes can differ (Lambrecht 1994). The standard way to demonstrate these differences are pairs of question-answers, such as in (i) below.

(i) a. A. - Who ran into the forest?  
    B. - Into the forest ran ROBIN HOOD.  
    B'.- ROBIN HOOD ran into the forest.  
(Rochemont 1986: 111-112)

b. A. - Where did Robin Hood1 run?  
    B. - Into the FOREST ran Robin Hood1.  
    B'.- *Robin Hood1 ran into the FOREST.
(3) A. a- Has John1 eaten [the sandwiches with mayo]2?
b- Hasn't John1 eaten [the sandwiches with mayo]2?
c- What has John1 done with [the sandwiches with mayo]2?
d- John1 is ill because of [the sandwiches with mayo]2, right? What's happened?

BEN. a- Yes.
b- Yes, he1 has. c- He1’s eaten them2. d- He1’s eaten them2.

Speaker BEN can answer each of the questions of Speaker A with respective sentences in (3B EN.a-d), i.e., (3B EN.a) is a felicitous answer of (3A.a), (3B EN.b) is so as regards (3A.b), and so on. Note that (3B EN.c) and (3B EN.d) are identical, although they are appropriate to answer different questions. Speakers BBP and BRU can properly answer RU and BP equivalent questions (3A.a), (3A.b) and (3A.c) by uttering the bare verb — comeu/съел “eat-pp”. This utterance is not an acceptable answer for question (3A.d). The reasons for this unacceptability will not be explored in the paper. Example (3) shows the variation of forms for the same proposition according to different questions.

3. State of the art

For tackling problem (i) in section 1, a number of theoretical and computational proposals to intersentential anaphora have been proposed (see Mitkov, 2002). A few works come to grips with the fact that subjectless sentences are challenging for SMTs. The gist of them is that a purely SMT is unable to deal with different pronominal systems, such as those of (non-)pro-drop languages (Gojun, 2010). Taku et al. (2015) mention that pro-drop languages can roughly be classified into two groups: consistent pro-drop (such as, Italian, Spanish and Czech) and discourse-related pro-drop languages (such as, Japanese and Chinese). According to them, most Romance and Balto-Slavic languages are consistently pro-drop. Biberauer et al. (2010) further discuss this classification and put forth a new taxonomy, adding a category called partial pro-drop (BP and RU). Most works that address different subject systems in SMT are focused on consistent subject languages and non-pro-drop languages.

Nakaiwa and Ikehara (1992) developed a system of correspondences between English and Japanese verbs and their relations with the respective argument structures. Besides this alignment system, they propose a “topicalized unit sentence” component, which contains nominal phrases that can serve as referents of anaphors. Their system can be applied to some extension to translations in mismatched (non-)pro-drop pairs of languages, but would not handle subjectless vYNAs, because the referents in the English output are not overt anaphoric elements (as in B ENa) and the verbs are not synonymous nor present the same semantic relation with their subject (as in B ENb). Goldwater & McClosky (2005) propose a system with pseudo-words to simulate correspondences between subjectless Czech sentences and English subject-verb sequences. Clearly, the case of vYNAs is not a matter of this strict one-to-one correspondence. Even if null constituents are projected on an abstract level, the number of (pseudo)words in the input and the output of vYNA contexts is not the same. Working on Italian-English translations, Gojun (2010) proposes a huge machinery to deal with the differences between the pro-drop Italian sentences and the non-pro-drop English outputs and vice-versa. Her solution is basically improving the alignment between Italian and English sentences, so that the Italian verbs correspond to the subject pronoun and verbs in English. Beyond its cost, this proposal does not solve the question of subjectless vYNAs in BP/RU translated into EN. Looking back at the example in (3), this model would predict the correspondence between (B BP/RU) and (B ENc), but would not deal with (B BP/RU) and (B ENa-b). Peral & Ferrández (2003) propose to tackle the difference between Spanish and English by analyzing the input at several steps by means of an “interlingua” approach. This enriched input is analyzed by a component that identifies and resolves anaphora. The authors mention successful results concerning anaphora identification and generation. However, they conclude that their system requires a further task which is to decide whether a pronoun must be realized.

In (i), the same propositional content is encoded by both the utterances B and B’, but not all forms are appropriate to answer the different questions. Standardly languages use intonational patterns to encode these differences, especially focal stress (here encoded into capital letters) and deaccenting. However, languages can non-rarely resort to other mechanisms to make the discourse structure clear, such as word order, active-passive transformation, pronominalization, ellipsis, null constituents, etc.
Such a task is crucial to cope with the possibilities of translation of vYNAs though. Le Nagard & Kehn (2010), Hardmeier & Federico (2010) and Loáciga (2013) also argue for considering some constraints for anaphor resolution in translating personal pronouns between English and French, German and English, French and Spanish, respectively. These approaches could generate grammatically acceptable sentences, but they would not correspond to the expected answers: a reply with all the anaphors in English would be more than minimally necessary to confirm or deny a direct question. Le Nagard & Kehn (2010) and Hardmeier & Federico (2010) advance though reasons for improving SMTs with cross-sentence dependencies and discourse structure.

Most rule-based proposals to improve SMT do not include an information structure component into their machinery. Notable exceptions in the literature about MT are Asher et al. (2004), Song & Bender (2011) and Guzmán et al. (2014). Guzmán et al. (2014) propose to bring together (i) a more structured linguistic information and (ii) an analysis beyond the sentence level. They claim that coherence relations must be incorporated into SMTs, by parsing and modeling Discourse Trees (cf. SDRT, see Asher & Lascarides 2003, *inter alia*). Although this approach is nascent and seems attractive, it is also costly and needs a strong development and establishment of theoretical and analytical categories before large-scale implementation. Moreover, in a series of papers, Kehler & Rohde (2013, 2014, *inter alia*) examine the role of rhetorical relations in anaphora resolution and production by means of psycholinguistic experiments. One of their main conclusions is that despite relevant in coreference interpretation, rhetorical structure has little (if any) significant importance in pronoun production, which is claimed by the authors to be related to information structure. Finally, it is worth mentioning the algorithm proposed by Taku et al. (2015) to deal with multiple languages, which consists of a mechanism of (i) null subject identification and (ii) null subject type estimation. They claim that task (i) need a global sentence analyzer, which can output the relation between predicate and arguments in a sentence, but task (ii) would not need anything else than the local context around the verb. Their algorithm is applied to subjectless outputs in English, so that it can generate grammatical sentences, but not simple yes or no answers or tag-answers.

4. Proposal

In the previous section, I have summarized some approaches to improve translations of pro-drop into non-pro-drop languages. They could provide a leap in the results shown in section 2. Most of them are, however, concerned primarily with syntactic and semantic correspondence. That is, as mentioned in section 2, allosentences would be statistically paired in the same way they are in the baseline, *i.e.* a standard SMT, and no significant improvement concerning pragmatics would be achieved. In this section, I present my proposal: to incorporate into SMTs a rule-based annotation component that relies on the relation between sentences, mainly information structure annotation. My proposal in this paper is closer to and build on Song & Bender (2011) and mainly on Ginzburg (2011). I thus assume three basic principles in the cross-sentencial relation: (i) coherence – questions are followed by answers which can raise new questions; (ii) conciseness – conversation is a highly efficient medium; and (iii) radical context dependency – isolated from their occurrence in a given context many utterances lose most of their importance. These principles are mostly uncontroversial. However, as pointed out by Ginzburg (2011, p. 2), what is at issue is whether such a relation is built into a grammar. Ginzburg (2011) develops a formal approach within the HPSG theory to deal with intersentential relations under the framework of Question under Discussion[QUD] (see Roberts, 1995, Buring, 1999). Song and Bender (2011) also use the HPSG theory, but their approach is embedded in many intra-theoretical assumptions that I am not able to detail here. Furthermore, Song and Bender (2011) are committed with standard categories in the level of information structure analysis, such as Topic, Focus, etc. As Roberts (1995) pointed out, these notions are built on the assumption that information structure is a sentence level. However, besides being notions whose definition is highly controversial, they are not able to fully account for “the range of kinds of contexts in which a given utterance is actually felicitous” (Roberts, 1995, p. 2). Moreover, most of these notions are diagnosed by question-answer pairs. My analysis empty such notions and directly bank only on the question-answer pair itself and some logical formalism to generalize over pairs.

Mainly I propose that the sentences are at first information-structurally analyzed, as in the sample in Fig. 3 below, and at last paired and statistically quantified. In this sample, the verb in BP or RU is tagged with three different underlying questions that it is able to answer – the questions under discussion QUD –, *cf.* example (3) in section 2. These three possibilities are paired with the same underlying questions and their correspondent answers in EN. In this pairing system, I presuppose that the underlying questions are somehow universal, following Roberts (1995), Buring (1999) and Ginzburg (2011), among others. Within this pairing system, no ungrammatical sentence is possible, since
the verb is not paired with a bare verb in the right side even though it is alone in the left side. In Fig. 3, I encode the presupposition given by the question as information that is to the left of | symbol. \( P \) stands for an open proposition inherited from the question, whose \( x \) is the focused element that satisfies and closes it (following Rooth, 1985, 1993, Roberts, 1995, *inter alia*). The matching QUDs are analyzed in parallel with the standard lexical pairing. If a verb tagged as *default* is lexically aligned with a verb with a full specified QUD, this is not taken as a valid correlation. The contexts in which these bare verbs are found are so discursively restricted and well-defined that the QUD can take precedence over lexical frequency without over-generating yes-no answers.

5. Methodology

I apply the proposal described in the previous section to a sample of a corpus of paired dialogues (436 vYNA-tokens from OPUS project, Tiedemann, 2012). I have selected the examples from the original version of 5 movies in each language and annotated them following the types proposed in Fig. 3. After having this input tagged, I have generated sentences in target language whose annotation is the same. Finally, I compared the set of sentences generated and the baseline provided by a standard SMT to the translation provided by the alignment in OPUS corpora, which is built on a time-based pairing between sentences (Tiedemann, 2012). The use of corpora based upon movie subtitles is not unheard of in the literature about linguistics. Many works have been using them to pursue monolingual studies (New et al. 2007, 2009) and bilingual parallel studies (Hardmeier, & Volk, 2009, etc.). New et al. (2007, 2009) use these texts to estimate frequency of words in spoken texts. They claim that (i) this kind of corpus is the closest possible to natural oral recorded texts (whose one hour of transcription takes around 40 hours for to be carried out) and (ii) word frequencies obtained from these corpora correlate highly with those from well-established sources such as Celex and Lexique for English and French. Word frequencies are highly responsive to corpora sources (Coady, & Aslin, 2003). Kilgarriff (1997) argues for using word frequencies to measure similarity between corpora. If so and New et al. (2007, 2009)'s claims are correct, movie subtitles are a reliable source of linguistic data. If a phenomenon closely linked to difference in corpora is claimed to be homogeneous across movie subtitles and other well-established corpora, a grammatical phenomenon such as the vYNAs is likely to be comparable in several corpus sources and movie-subtitle corpora. Movie subtitles are the nearest source of dialogue-like written texts that are available in large amounts. As a basic instance of communication, the importance of dialogues is undeniable. The study of movie subtitle corpora may cast light on relation among utterances built into a grammar, because they are less likely to present performance effects that blur the analysis of natural conversation. The corpus studied here – OPUS – is probably the largest free corpus and is constantly growing. Data comes from several domains. In the paper, subcorpora, called OpenSubtitles (http://www.opensubtitles.org/), was taken to be the sample because it is constituted primarily of paired subtitle dialogues (Lison, & Tiedemann, 2016).

![Fig. 3. A simplified version of discursively tagged and paired sample of sentences.](image)

6. Results

The analysis was carried out with a comparing table akin to Table 3 below (two examples in BP and two in RU).

<table>
<thead>
<tr>
<th></th>
<th>input_question</th>
<th>input_answer</th>
<th>output_ans_b</th>
<th>baseline</th>
<th>Syn_base</th>
<th>sem_base</th>
<th>inf_struct_base</th>
<th>paired_question</th>
<th>paired_answer</th>
<th>prediction</th>
<th>matching</th>
<th>overall_eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chegou cobertura?</td>
<td>Chegou.</td>
<td>Has arrived.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Did the reinforcement arrive?</td>
<td>Yes.</td>
<td>Yes.</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. A sample of analysis from the corpora study.
The corpora and the statistical analysis confirm that the discourse tagging improves SMTs. In Table 3, the lowest raw shows a descriptive analysis. The source language (BP and RU) were inserted in the baseline SMT and the output was analyzed in a binary notation for grammatical acceptability, semantic similarity and information structure adequacy. The output of the model proposed was analyzed in terms of similarity with the paired text and an overall evaluation (taking into consideration semantics and pragmatics). As shown in Table 3, the improvement is relevant. However, this is a descriptive analysis. Table 4 below shows a binomial logistic regression carried out with the same data, but evaluated in terms of missing arguments, incorrect subject type, correct verbal (item and agreement), particle presence (yes or no), tag answer utilization, full sentence output and non-sentential “other” elements. The second and forth columns present the average of the baseline and my model (called information structure-based [ISB]SMT). The third and fifth columns show the significance of the difference of accuracy of the baseline and the ISBMT as regards the comparison with the human paired text. Finally the right-most column show the results of testing the hypothesis H0: Ι= human translation that contains only an intercept term. A likelihood ratio test comparing the baseline and ISBMT models was performed using the anova() function with the test="Chisq".

<table>
<thead>
<tr>
<th>parameters</th>
<th>baseline</th>
<th>difference_significance</th>
<th>ISBMT</th>
<th>difference_significance</th>
<th>human_paired_text</th>
<th>model_comparison_P(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing_arguments</td>
<td>0.5948</td>
<td>*</td>
<td>0.9971</td>
<td>n.s.</td>
<td>0.9741</td>
<td>0.002114 **/n.s.</td>
</tr>
<tr>
<td>incorrect_subject</td>
<td>0.6513</td>
<td>***</td>
<td>0.4286</td>
<td>n.s.</td>
<td>0.3824</td>
<td>&lt; 2.2e-16 ***/n.s.</td>
</tr>
<tr>
<td>correct_verb</td>
<td>0.4463</td>
<td>***</td>
<td>0.5626</td>
<td>n.s.</td>
<td>0.5331</td>
<td>&lt; 2.2e-16 ***/n.s.</td>
</tr>
<tr>
<td>particle</td>
<td>0.4352</td>
<td>***</td>
<td>0.5575</td>
<td>n.s.</td>
<td>0.5401</td>
<td>&lt; 2.2e-16 ***/n.s.</td>
</tr>
<tr>
<td>tag_answer</td>
<td>0.06322</td>
<td>*</td>
<td>0.07671</td>
<td>n.s.</td>
<td>0.07759</td>
<td>0.02243 */n.s.</td>
</tr>
<tr>
<td>full_sentence</td>
<td>0.6121</td>
<td>***</td>
<td>0.4971</td>
<td>n.s.</td>
<td>0.5029</td>
<td>&lt; 2.2e-16 ***/n.s.</td>
</tr>
<tr>
<td>other</td>
<td>0.07184</td>
<td>***</td>
<td>0.1063</td>
<td>n.s.</td>
<td>0.1121</td>
<td>3.319e-16 ***/n.s.</td>
</tr>
</tbody>
</table>

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘n.s.’ 1

7. Discussion

SMTs must be evaluated in terms of many aspects of translation, including fidelity, adequacy, grammaticality and fluency of the output. When the range of variation concerning answers to yes-no questions is taken into consideration, SMTs fail in most of these criteria. In this paper, I showed that translations of vYNAs in BP and RU are challenging for SMTs. Furthermore, I have put forth a pragmatic-based tagging system to improve the statistical correspondence and further develop SMT. Besides generating grammatically acceptable sentences, the results shown here contribute to improve semantic-pragmatic precision in translations, since the relations on such levels are preserved. I do not claim throughout this paper that this solution goes beyond answers yes-no questions. However, the QUd approach to deal with underlying information structure appears fruitful and can be extended to other contexts in which semantics and
pragmatics play a role in translation.

Acknowledgements

I would like to thank Barbara Hemforth and Philip Miller for the comments in parts of this ongoing work; Maria Eiferova, for the data and constant inspiration; and the CAPES Foundation for the financial support.

References