Machine Translation of Non-Contiguous Multiword Units

Anabela Barreiro[1] **and Fernando Batista**[1,2]
(1) INESC-ID Lisboa, Portugal
(2) ISCTE-IUL, Instituto Universitário de Lisboa, Portugal
{anabela.barreiro, fernando.batista}@inesc-id.pt

Abstract

Non-adjacent linguistic phenomena such as non-contiguous multiwords and other phrasal units containing insertions, i.e., words that are not part of the unit, are difficult to process and remain a problem for NLP applications. Non-contiguous multiword units are common across languages and constitute some of the most important challenges to high quality machine translation. This paper presents an empirical analysis of non-contiguous multiwords, and highlights our use of the Logos Model and the Semtab function to deploy semantic knowledge to align non-contiguous multiword units with the goal to translate these units with high fidelity. The phrase level manual alignments illustrated in the paper were produced with the CLUE-Aligner, a Cross-Language Unit Elicitation alignment tool.

1 Introduction

Recently, in natural language processing (NLP), there has been an increasing interest in multiword units and in the problems they raise. Multiword units, most commonly known as multiword expressions[1], have been defined by Baldwin and Kim (2010) as "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". Compositionality is the property that makes the automatic processing of multiword units particularly challenging. Multiword units occur very frequently with different degrees of compositionality. Some represent free combinations, such as the English noun phrase *round table*, (i.e., *meeting*), some have opaque meanings, where the meaning of the unit cannot be deduced from the meaning of its individual constituents, such as *piece of cake* used figuratively with the meaning of *something easy to do*, or *pay a visit* equivalent to the verb *visit*. Translations of multiword units are often idiomatic and unpredictable and a word-for-word translation may result in poor quality translation (*acid test*). Additionally, in many cases, the idiom does not exist, or exists with a different structural and lexical form, in the target language (*raining cats and dogs*). Finally, the morpho-syntactic properties of multiword units allow, in some cases, the insertion of external elements into the unit (*go for a* [INSERTION] *ride*).

Notwithstanding the efforts undertaken to improve multiword unit processing, lack of formalization still triggers problems with the syntactic and semantic analysis of sentences where multiwords occur and impairs the performance of NLP systems, affecting especially machine translation (MT). Whenever a multiword unit contains insertions, there is a remote dependency that contributes to additional difficulties in analysing and translating that multiword unit. The causes for poor quality translation of lexical and semantico-syntactic phenomena, namely cross-linguistic multiwords and other phrasal units, can be summed up in three points:

1. Current methodologies rely mostly on statistical techniques to train and evaluate MT systems. Statistical machine translation (SMT) models are built on the grounds of alignment

[1] This term has also been designated *inter alia* as "multiword lexical itens", "phraseological units" and "fixed expressions", with slight variations in scope and meaning.

Proceedings of DiscoNLP 2016, pages 22–30,
San Diego, California, June 17, 2016. ©2016 Association for Computational Linguistics

pairs acquired mostly automatically. Unsupervised learning approaches adopted by SMT systems use probabilistic alignments where linguistic knowledge is still limited. Inability to identify multiword units correctly often results in translation deficiencies.

2. Shortcomings in current state of the art supervised learning and manual word alignment standard practices, such as lack of publicly available manual multilingual datasets, and lack of linguistically motivated alignment guidelines impose significant constraints on translation quality, because they disregard non-adjacent linguistic phenomena or syntactic discontinuity.

3. Current tools are incapable of assisting, in an efficient way, human annotators in the task of identifying correctly non-contiguous multiwords and produce rules from them.

This paper discusses the aforementioned problems from an empirical point of view and provides a solution for them in an experimental research inspired in the Logos Model to machine translation (Scott, 2003; Barreiro et al., 2011).[2] We use the Europarl corpus (Koehn, 2005) to illustrate the kind of linguistic knowledge that needs to be represented in future alignment tasks, with a special focus on the alignment challenges presented by non-contiguous multiword units. The alignment examples in the paper were annotated with the CLUE-Aligner tool (Barreiro et al., 2016). Even though similar in name to the "clue alignment approach" (Tiedemann, 2003; Tiedemann, 2011), mainly devoted to word-level alignment, our approach is theoretically and methodologically different with a focus on phrase alignment, contemplating multiwords and linguistically-relevant phrasal units. In addition, in our approach, CLUE is an acronym for "**C**ross-**L**anguage **U**nit **E**licitation", where a source and a target language can correspond to the same language as in the case of paraphrases.

2 Addressing the Challenges

The first problem highlighted in section 1, has been addressed by the creation of an empirical basis for identifying support verb constructions in a corpus and understand the impact of these multiword units on translation quality. Barreiro et al. (2014) evaluated support verb constructions by two MT systems, OpenLogos and Google Translate, and concluded that neither of these systems translates them well. Overall, OpenLogos suffers from a weak lexicon, while Google Translate translation errors are more of a structural nature. Although the translations are still problematic, the Logos Model presents an advantage with regards to the SMT approach: the Logos Model relies on deep semantico-syntactic analysis to translate not only contiguous multiword units, such as the support verb construction *to draw a distinction between*, but also non-contiguous multiword units, such as the support verb construction *to bring* [INSERTION] *to a conclusion*.

The second problem reported has been approached by the creation of manually annotated alignments – the Gold CLUE4Translation – which represent an important asset in the development of MT systems. Supervised learning uses manual alignments and aims at taking context, syntax and other grammatical and semantic knowledge into consideration. The Logos Model served as inspiration to deploy this linguistic knowledge into the alignment task through the identification of translation relationships among words, multiwords or phrasal units in bilingual parallel sentences, i.e., sentence pairs that have been identified as translation of each other. It also inspired the establishment of new linguistically-motivated alignment guidelines for pairs of translation units – the CLUE4Translation Alignment Guidelines – that aim to improve the quality of the (machine) translation of multiword units, among other linguistic phenomena.

The third problem was tackled by the creation of a solution for the annotation of non-contiguous multiwords and other phrasal units – the aforementioned CLUE-Aligner[3] – a web alignment tool that places a special emphasis in the annotation of pairs of semantically equivalent non-adjacent structures in mono-

[2] The Logos Model underlies both the commercial system and its degraded open source version OpenLogos.

[3] https://esperto.l2f.inesc-id.pt/esperto/aligner/index.pl?

lingual and bilingual parallel sentences. The pairs of non-contiguous multiword units and phrasal expressions can be used in rule development.

3 The Logos Model

The struggles of SMT with multiwords have been reported in several research works (Barreiro et al., 2013; Kordoni and Simova, 2014; Barreiro et al., 2014; Semmar, 2012), among others. Multiword units are a source of mistranslations not only by MT systems, but also by professional translators, in part because they are a source of various contextual nuances, but also because they can be non-contiguous.

For example, verbal expressions such as the English prepositional verb *to deal with* take difference senses (and translations) depending on contexts, typically their object or prepositional phrase complement. If the context of the verb is *to deal with questions*, as in example (1), then the French translation should be *s'occuper de* (*to be busy with*). On the other hand, if the context is *he proved unable to deal with the problem*, then the translation should be the translation of its paraphrase *handle the problem*. However, if the context is *he refused to deal with the problem*, then the translation would be a translation of the paraphrase *analyse and try to solve the problem*. These different nuances are related to the ambiguity and weakness of the verb *deal* and the different meanings of the predicate-like nouns *questions* (*issues*, *topics*, *interrogations*, etc.) or *problem* (*difficulty*, *exercise*, etc.). It is the meaning of these nouns that triggers the different translations of *deal*, just like the verb *take* will have different translations depending on the predicate noun it supports (*walk*, *responsibility*, *comfort*, etc.). Therefore, the two slightly different meanings for *problem* in the last two examples explain the distinct paraphrase: *handle* in one case, and *analyze and try to solve* in the other case.

In the Europarl corpus used in our exploratory study not all translations are optimal and often translational equivalents are approximate rather than exact. Therefore, the English prepositional verb *to deal with* in example (1) is translated in the Romance languages as *dedicarse a* (*engage in*) in Spanish, the reflexive *s'attacher a* (*focus on/stick to*) in French, and *centrar-se em* (*concentrate/center* (*their thoughts*) *on*) in Portuguese. The different translations of *deal* are related to the idiomatic ways that predicate nouns select their support verbs in different languages: *take a vow* in English, but '*make a vow*' in the Romance languages (*hacer* in Spanish, *faire* in French, and *fazer* in Portuguese).

(1) $_{EN}$ - our Asian partners prefer **to deal with** questions which unite us

$_{ES}$ - nuestros socios asiáticos prefieren **dedicarse a** las questiones que nos unen

$_{FR}$ - nos partenaires asiatiques préfèrent **s'attacher à** ([**a**+a]) ce qui nous unit

$_{PT}$ - os nossos parceiros asiáticos preferem **centrar-se** unicamente **n**as ([**em**+as]) questões comuns

If the different nuances of a verbal expression are difficult to capture even for translators, it is not surprising that these expressions are poorly translated by MT systems, unless these systems integrate semantic or contextual knowledge and apply it to the translation process, as illustrated in example (2). The French MT output of example (1) by the Google Translate (GT) system is a literal translation where no context has been taken into consideration. However, the OpenLogos translation is correct and even of a higher quality than that provided by a professional translator in the Europarl corpus (*s'occuper de* is more precise than *s'attacher a* in that context).

(2) $_{FR-GT}$ - nos partenaires asiatiques préfèrent ***traiter avec des** questions qui nous unissent

$_{FR-OL}$ - nos associés asiatiques préfèrent **s'occuper des** questions qui nous unissent

The precision in the OpenLogos translation is associated with the application of a Semtab contextual pattern-rule, which is a deep structure pattern that matches on/applies to a great variety of surface structures:

(3) deal(VI) with N(questions) = s'occuper de N[4]

This Semtab pattern-rule states that, when followed by the direct object noun *questions* or a noun of the same semantico-syntactic class, the verb is translated as *s'occuper de*, overriding the default

[4] Here we only display the comment line of the Semtab rule, not the rule itself or what it does in terms of the Logos language. The rule notation is arcane due to its numeric representation and it would take a larger effort to explain the use and meaning of the distinct codes in the Logos Model.

Pattern	#occurences	#unique
bring [ ] *to a conclusion*	114	84
set [ ] *in motion*	162	140
play [ ] *role*	5165	1216
take [ ] *interest in*	360	163
keep [ ] *informed about*	77	58

Table 1: Statistics from a subset of Europarl.

dictionary translation for this verb. The power of this rule is that it allows the translation system to recognize and analyze multiword units, even when the elements of the multiword units are non-contiguous. The alignment of multiword units to feed a SMT system needs to reflect these semantic nuances, in a similar way to the way the Logos Model uses data-driven pattern-rules to account for these nuances.[5] This proves that alignments that mirror Semtab semantic and contextual pattern-rules of the Logos Model can help create new MT systems and improve existing ones.

4 Alignment of Non-Contiguous Multiwords Inspired by Logos

Non-contiguous multiword units are difficult to recognize and process causing many MT systems to fail in providing the correct translations. For SMT systems, non-contiguous multiword units represent a significant challenge to a correct word and phrase alignment (Shen et al., 2009).

The Europarl corpus contains a significant number of occurrences of non-contiguous multiword units, such as the support verb constructions illustrated in subsections 4.1 – 4.5, which are a source of translation errors due to incorrect alignment. Table 1 shows the number of occurrences for five different cases of non-contiguous multiword units in a subset of the Europarl corpus, containing about 47.4 million words, where the search was performed using all forms of each verb. The third pattern is the most common type of multiword unit and also the one with the highest spectrum of common usages. About 83% of the forms occur more than once. The fourth pattern occurs more than once 65% of the times, revealing these commonly adopted constructions appear in many different forms. On the other hand, the first, second and fifth patterns occur only once, 62%, 78% and 62% of the times, thus suggesting that learning automatic models to deal with these type of constructions may not be straightforward.

The remainder of this section discusses each case taking into account the Logos Model and showing how each alignment is represented in CLUE-Aligner.

4.1 *bring* [ ] *to a conclusion*

In example (4), the English non-contiguous support verb construction *bring* [INSERTION] *to a conclusion* places the predicate noun *conclusion*, with its adnominal modifiers, ten words apart from the support verb *bring*. The Spanish, French, and Portuguese translation equivalents adopt different stylistic variants and simpler surface structures (i.e., syntax) by transforming the support verb construction into semantically equivalent verbal constructions, the single verb *acelerar* (*speed up, accelerate*) in Spanish or the compound verbs*faire avancer* (*make advance*) and *apressar-se a apresentar* (*hurry to present*) in French and Portuguese, respectively.

(4) *EN* - I would urge the European Commission to **bring** the process of adopting the directive to on additional pensions **to a conclusion**

ES - insto a la comisión europea para que **acelere** la directiva sobre pensiones complementares

FR - j'insiste auprès de la comission européenne pour **faire avancer** la directive sur les pensions complémentaires

PT - exorto a comissão europeia a **apressar-se a apresentar** a directiva relativa as pensões complementares

This non-contiguous support verb construction, with a remote placement of one of the components of the unit, represents a difficult unit to align and to translate. In general, statistical (or statistically-based) MT systems translate fairly well contiguous multiword units taking into account context (surrounding word strings). However, purely statistical phrase-based MT systems translate poorly multiwords that contain elements placed remotely

[5] Unlike rule-based MT models, the Logos Model is not rule-driven, but data-driven, i.e., in Logos, patterns, not rules, do the matching. So, in Logos, a rule refers to the action component once a match is made. Like SMT, it is possible to apply the same techniques to the data, which in the Logos Model is not literal words but semantico-syntactic (SAL) patterns or entities. This is the reason why it makes sense to train a machine learning system to learn new SAL patterns based on alignments, instead of on the conventionally used SMT patterns.

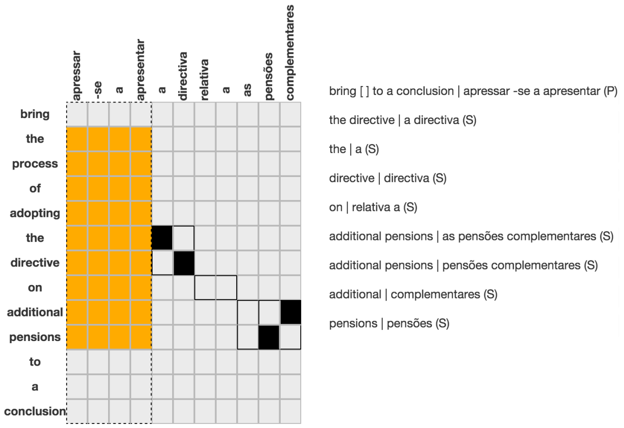
apressar
-se
a
apresentar
a
directiva
relativa
a
as
pensões
complementares
bring
the
process
of
adopting
the
directive
on
additional
pensions
to
a
conclusion
bring [ ] to a conclusion | apressar -se a apresentar (P)
the directive | a directiva (S)
the | a (S)
directive | directiva (S)
on | relativa a (S)
additional pensions | as pensões complementares (S)
additional pensions | pensões complementares (S)
additional | complementares (S)
pensions | pensões (S)

Figure 1: Alignment of *bring* [ ] *to a conclusion*

(long distance dependency), as illustrated in the Portuguese translation of the support verb construction by Google Translate in example (5), where the verb is missing.

(5) *PT−GT* - Gostaria de exortar a Comissão Europeia a que o processo de adopção da directiva para as pensões adicionais ***para** uma conclusão.

Figure 1 represents the P-alignment of the non-contiguous support verb construction *bring* [ ] *to a conclusion* with its contiguous equivalent compound verb *apressar-se a apresentar* in Portuguese.

The most common expression found in our subset of the Europarl corpus is *bring this matter to a conclusion* and the remaining expressions occur very few times.

4.2 *setting* [ ] *in motion*

Often in translation, a non-contiguous expression in a source language can be maintained in the target language or replaced by an equivalent but contiguous expression that conveys the same meaning. It can also be transformed into a simpler contiguous syntactic structure, such as a single word. For example, the Portuguese translation for the non-contiguous English support verb construction *set in motion* in example (6) is the single verb *empreender* (*undertake*). Both Spanish and French maintain the support verb constructions (*llevar a cabo* and *mettre en chantier*), but they are contiguous, having no insertions. The presence of a non-contiguous expression in one of the sentences of the language pair causes additional complexity to the alignment task, which we are able to solve with the Logos Model approach.

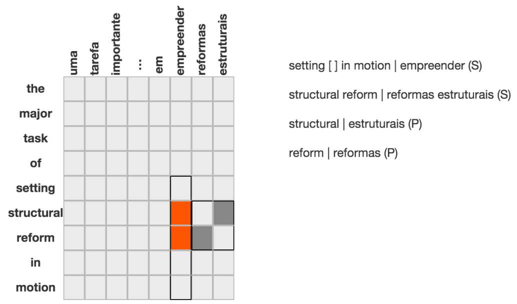

Figure 2: Alignment of *setting* [ ] *in motion*

(6) *EN* - many member states thus have the major task of **setting** structural reform **in motion**

ES - he aquí por lo tanto una tarea de gran importancia para que numerosos estados miembros **lleven a cabo** reformas estructurales

FR - il y a donc là une táche considérable pour beaucoup d'états membres, celle de **mettre en chantier** des réformes structurelles

PT - há, portanto, uma tarefa importante para muitos estados-membros em **empreender** reformas estruturais

Figure 2 represents the alignment of the non-contiguous support verb construction *setting* [ ] *in motion* with the single verb *empreender* in Portuguese.

4.3 *play* [ ] *role*

In some cases, the verbal expression is always expressed in the form of a support verb construction, which is the case of *play* [INSERTION] *role*, because there is no semantically equivalent single verb. The support verb can take several forms, i.e., the construction can be stylistically different (Barreiro, 2009). Figure 3 exemplifies the adjective modifier insertions *increasingly predominant* in the English sentence. These insertions are excluded from the English–Portuguese alignment pair *play* [ ] *a role* – *desempenham um papel* and aligned separately.

4.4 *take* [ ] *interest in*

Non-contiguous prepositional verbs are aligned together with the preposition. Example (7) illustrates the alignment of the English support verb construction *take* [ ] *interest in* with its semantically equivalent prepositional verbs in the Romance languages. In this support verb construction, the preposition *in*

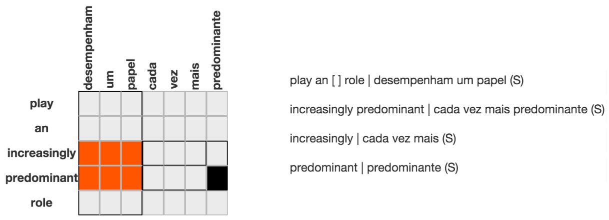
desempenham
um
papel
cada
vez
mais
predominante
play
an
increasingly
predominant
role
play an [ ] role | desempenham um papel (S)
increasingly predominant | cada vez mais predominante (S)
increasingly | cada vez mais (S)
predominant | predominante (S)

Figure 3: Alignment of *play* [ ] *role*

is selected by the predicate noun *interest*, and not by the support verb *take*. In the Romance languages, the prepositions are selected by strong verbs: *ocuparse* [ ] *de* in Spanish, *s'occuper* [ ] *des* in French, and *debruçar-se* [ ] *sobre* in Portuguese. The adjectival insertion *special* aligns with the adverbial insertions in the Romance languages: *en especial* in Spanish, *en particulier* in French, and *em especial* in Portuguese.

(7) *EN* - the committee on employment and social affairs **took a** special **interest in** types of supplementary pension funds

ES - la comisión de empleo y de asuntos sociales **se ha ocupado** en especial **de** las modalidades de la asistencia suplementaria a la tercera edade

FR - la commission de l'emploi et des affaires sociales **s'est** en particulier **occupée d**es ([**de**+les]) différentes formes de retraite complémentaire

PT - a comissão do emprego e dos assuntos sociais **debruçou-se** em especial **sobre** as possibilidades existentes para regimes complementares de reforma

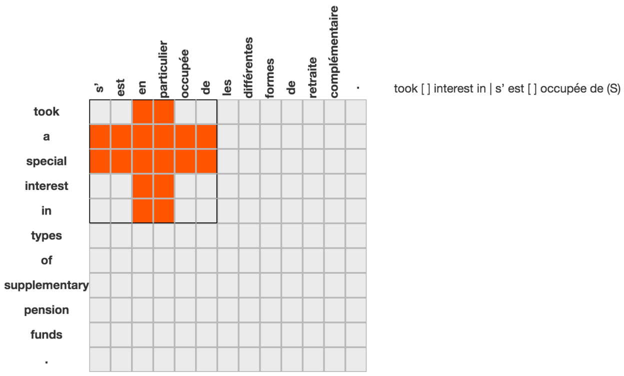

Figure 4: Alignment of *took* [ ] *interest in*

Figure 4 represents the alignment of the non-contiguous prepositional verb *took* [ ] *interest in* with the corresponding reflexive prepositional verb *s'occuper de* in French.

4.5 *keep* [ ] *informed about*

Prepositional adjectives align as internal elements of support verb constructions. Example (8) illustrates the alignment of contiguous prepositional adjectives, with the exception of Spanish. The prepositional adjective in Spanish contains an adverbial insertion (*periódicamente*) between the adjective *informados* and the preposition *de*. In English, French, and Portuguese, the adverbs occur before the prepositional adjectives. Therefore, the contiguous prepositional adjective *informed about* in English aligns with its semantically equivalent prepositional adjectives *informés des* in French, and *informados acerca d*(*os*) in Portuguese.

(8) *EN* - calling on the commission **to keep us** regularly **informed about** recent developments

ES - pidiendo a la comisión que **nos mantenga informados** periódicamente **de** lo que vaya ocurriendo

FR - appelant la commission à **nous tenir** régulièrement **informés des** ([**de**+les]) derniers développements de ce dossier

PT - em que se apela à comissão para que **nos mantenha** regularmente **informados acerca d**os ([**de**+os]) progressos que se forem realizando

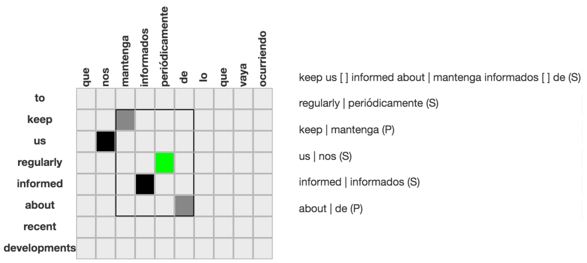

Figure 5: Alignment of *keep* [ ] *informed about*

Figure 4 represents the alignment of the non-contiguous support verb construction with a prepositional adjective *keep* [ ] *informed about* with its Spanish equivalent *mantenga informados de*.

5 Advantages of the Logos Model

Former word alignment techniques, even when they contemplated multiword unit alignments, were unable to present a consistent and efficient solution to process non-contiguous expressions. The advantage of the Logos Model with regards to non-contiguous multiword units is its ability to relate constituents that are apart (even very far apart) in the sentence. Semtab is an effective way of analysing and translating words in context, especially when the context is remote. In addition to this, Semtab also allows generalizing between alternative forms of the same multiword, phrase or expression. For example, it presents the possibility of generalizing translations of *take a walk* to translations of *walk*, if one of these two is found in the training corpus. Similarly, closed class items or highly frequent multiwords and phrases might be learnt quickly and be translated correctly by a SMT system, but open class items or less frequent multiwords and phrases might present more challenging problems that can be observed in MT translations, but also in non-native speakerisms, such as the choice of a support verb for a particular support verb construction (e.g., *make a visit* or *pay a visit*?), which can be robustely corrected by the use of Semtab.

Independently of the MT approach, the most important consideration with respect to multiword units is that they should never be processed on a word-for-word basis, because they represent atomic semantico-syntactic and translation units and cannot be broken down into constituent parts in any alignment process. Given that SMT translation quality depends on the quality of the alignments, it is necessary a better representation of multiword units, and greater amounts of training data. Only a more general representation and access to lexica will cause an impact on unseen multiwords. Therefore, linguistic knowledge "elicited" in the alignment process and the use of a more refined alignment tool can solve some of the problems related to multiword unit alignment, when it is so relevant that these alignments mirror the unity of the expression.

6 Analysis of Preliminary Results

Taking into account the search performed in section 4 and the corresponding results summarised in Table 1, we have analysed the first 20 sentences extracted from the subset corpus for each one of the multiword cases. In order to assess the current translation quality of each one of the previously described cases, we translated each sample using Google Translate and performed an empirical evaluation of the achieved results.

For the support verb construction *bring* [ ] *to a conclusion*, we categorized all 20 translations as incorrect, inadequate or non-optimal. Example (9) illustrates a literal, unnatural Portuguese translation *trazer a uma conclusão* for this support verb construction, where a paraphrase of it, such as *concluir* or *terminar este dossier*, represents a higher quality translation.

(9) $_{EN}$ - The Council is full of good intentions to do all it can **to bring** this dossier **to a conclusion**

$_{PT-GT}$ - O Conselho está cheio de boas intenções para fazer todo o possível para ***trazer** este dossier **a uma conclusão**

We also categorized all 20 translations of sentences with the support verb construction *set* [ ] *in motion* as incorrect. Example (10) illustrates a literal, incorrect translation *estabeleceu* [ ] *em movimento*, instead of *iniciou* or *pôs em marcha*.

(10) $_{EN}$ - it was the Polish Solidarnosc movement which **set** the downfall of the Soviet superpower **in motion** 20 years ago

$_{PT-GT}$ - foi o movimento Solidarnosc polonês que ***estabeleceu** a queda da superpotência soviética **em movimento** há 20 anos

For the support verb construction *play* [ ] *role*, we categorized 8 of the 20 translations as incorrect, inadequate or non-optimal. Example (11) illustrates a literal, incorrect translation *jogar o papel*, instead of *desempenhar o papel*.

(11) $_{EN}$ - the European Parliament is not prepared to simply **play the role of** observer

$_{PT-GT}$ - o Parlamento Europeu não está preparado para simplesmente ***jogar o papel de** observador

For the support verb construction *take* [ ] *interest in*, we categorized 16 of the 20 translations as incorrect, inadequate or non-optimal. Example (12) illustrates the consequences that an incorrect approach to non-contiguous support verb constructions (and other multiword units) have over translation, which is responsible for the incorrect agreement between noun (*interesse* is masculine) and adjective (*morna* is feminine), but also for the non-optimal translation of the support verb. A higher quality translation would use the non-elementary support verbs *manifeste* or *demonstre*, in the present subjunctive instead of in the infinitive form (unlike the incorrectly chosen support verb form *ter*).

(12) $_{EN}$ - It is unacceptable for the Commission only **to take** a lukewarm **interest in** a country

$_{PT-GT}$ - É inaceitável que a Comissão só a ***ter** um **interesse** morna **em** um país

For the support verb construction *keep informed about*, we categorized 9 of the 20 translations as incorrect, inadequate or non-optimal. Example (13) illustrates an incorrect translation *tem* [...] *manteve informados sobre* for this support verb construction, which should be translated as (*que nos*) *tem mantido informados*, or (*que nos*) *tem informado*, among others.

(13) $_{EN}$ - We have a Commissioner who has played and still is playing a major role in this enlargement, who **has** constantly **kept** us **informed about** what he was doing and with whom we have clearly always been on the same wavelength from a political point of view.

$_{PT-GT}$ - Temos um Comissário, que desempenhou e continua a desempenhar um papel importante no este alargamento, que ***tem** constantemente nos ***manteve informados sobre** o que ele estava fazendo e com quem temos claramente sido sempre no mesmo comprimento de onda de um ponto de vista político

In addition to the lexical problems related to the translation of non-contiguous multiword units, there are also structural errors, such as lack of agreement (e.g., *para nos manter regular e estreitamente* ***informado** *sobre*; *que o Parlamento* ***ser** *bem* ***informados** *sobre*) and incorrect word order (se conseguirmos ***a** adoptar e defini-**lo** em movimento), among others.

Even though, we have analysed just a few cases, the findings point to a general lack of quality in the translation of non-contiguous support verb constructions, which appear to be also true for other types of non-contiguous multiword units and phrasal expressions. A broader quantification of the phenomenon would help validating our preliminary results. A subsequent work could evaluate the performance of hierarchical phrase-based, syntax-based, and neural network translation models, which have the theoretical capacity to learn non-contiguous expressions.

7 Conclusions and Future Directions

This paper aims to prove that standard MT systems can benefit significantly by assuming a correct processing of non-contiguous multiword units, which current approaches are not exploring efficiently. The amount of post-editing effort can be reduced by increasing the quality of the alignments.

Non-contiguous support verb constructions processing, recognition and translation is a challenging problem when using alignment techniques. Some methodologies are inefficient in the sense that they violate the intrinsic property of the unit as an atomic group of elements when aligning them individually or when not respecting the correct boundaries of the unit.

Another problem concerns manual multilingual alignment scarcity and lack of linguistically rich alignment guidelines. Previously proposed word alignments guidelines cover cross-linguistic phenomena superficially, excluding the important alignment challenges (and challenges to machine translation) presented by non-contiguous support verb constructions and other multiwords and phrasal units.

This paper presented the reasons why non-contiguous correct and non-ambiguous alignment is important, and showed how alignment challenges have been addressed in the Logos Model. This model inspired us to create an alignment methodology that allows the correct alignment of non-contiguous multiwords and other phrasal units, which we were able to represent graphically in CLUE-Aligner, an alignment tool that handles non-adjacent structures in an appropriate way.

A strategic follow-up of this experimental research is to extract translation rules from manually annotated corpora and enhance an initially created Gold standard of manual annotations of bilingual alignment pairs to feed CLUE-Aligner, based on alignment decisions documented in the work in progress set of CLUE Alignment Guidelines. Therefore, future work aims the enhancement of CLUE-Aligner to align and extract automatically large amounts of alignment pairs to be applied to MT case studies, with an ultimate goal to improve translation applications. Linguistically-based alignments extracted from good quality translation corpora can contribute to increased precision and recall in SMT systems, with the subsequent improvement of translation quality. They are also a valuable asset for applications that require monolingual paraphrases.

This paper could not be ended without without underlining the great importance of paraphrases in the translation process. Future machine translation requires paraphrastic knowledge that allows to choose among possible translations, the best translation for a multiword unit or phrasal expression in the particular sentence where it occurs (i.e., in context), as illustrated in example (14).

(14) $_{EN}$ - It is time **to bring** this issue **to a conclusion**
$_{EN}$ - We must **bring** this episode **to a conclusion**

$_{PT}$ - Está na altura de **resolver** esta questão
$_{PT}$ - Chegou a hora de **concluir** este assunto
$_{PT}$ - **Ponhamos um ponto final** neste tema
$_{PT}$ - Temos de **concluir** este episódio.

Acknowledgements

This research work was supported by Fundação para a Ciência e Tecnologia (FCT), under reference UID/CEC/50021/2013, project eSPERTo – EXPL/MHC-LIN/2260/2013, and post-doctoral grant SFRH/BPD/91446/2012. The authors wish to thank Brigitte Orliac and Bud Scott for commenting on the portion of this paper related to the Logos Model.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

Anabela Barreiro, Bernard Scott, Walter Kasper, and Bernd Kiefer. 2011. OpenLogos Rule-Based Machine Translation: Philosophy, Model, Resources and Customization. *Machine Translation*, 25(2):107–126.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, and Fernando Batista. 2013. When Multiwords Go Bad in Machine Translation. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Machine Translation Summit XIV*.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, Susanne Preuss, Kutz Arrieta, Wang Ling, Fernando Batista, and Isabel Trancoso. 2014. Linguistic Evaluation of Support Verb Constructions by OpenLogos and Google Translate. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 35–40. ELRA.

Anabela Barreiro, Francisco Raposo, and Tiago Luís. 2016. CLUE-Aligner: An Alignment Tool to Annotate Pairs of Paraphrastic and Translation Units. In Nicoletta Calzolari et al., editor, *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, pages –. ELRA.

Anabela Barreiro. 2009. *Make it Simple with Paraphrases: Automated Paraphrasing for Authoring Aids and Machine Translation*. Ph.D. thesis, Universidade do Porto, Portugal.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Valia Kordoni and Iliana Simova. 2014. Multiword expressions in machine translation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. ELRA.

Bernard (Bud) Scott. 2003. The Logos Model: An Historical Perspective. *Machine Translation*, 18(1):1–72.

Dhouha Semmar. 2012. Identifying Bilingual Multi-Word Expressions for Statistical Machine Translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12)*, pages 23–25. ELRA.

Libin Shen, Jinxi Xu, Bing Zhang, Spyros Matsoukas, and Ralph Weischedel. 2009. Effective use of linguistic and contextual information for statistical machine translation. In *EMNLP 09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 72–80.

Jörg Tiedemann. 2003. Combining clues for word alignment. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 12–17, Budapest, Hungary.

Jörg Tiedemann. 2011. *Bitext Alignment*. Morgan and Claypool.