Dictionary of Multiword Expressions for Translation into Highly Inflected Languages

Daiga Deksne, Raivis Skadiņš, Inguna Skadiņa
Affiliation information
Tilde (Riga, Latvia)
E-mail: daiga.deksne@tilde.lv, raivis.skadins@ tilde.lv, inguna.skadina@ tilde.lv

Abstract

Treatment of Multiword Expressions (MWEs) is one of the most complicated issues in natural language processing, especially in Machine Translation (MT). The paper presents dictionary of MWEs for a English-Latvian MT system, demonstrating a way how MWEs could be handled for inflected languages with rich morphology and rather free word order. The proposed dictionary of MWEs consists of two constituents – a lexicon of phrases and a set of MWE rules. The lexicon of phrases is rather similar to translation lexicon of the MT system, while MWE rules describe syntactic structure of the source and target sentence allowing correct transformation of different MWE types into the target language and ensuring correct syntactic structure. The paper demonstrates this approach on different MWE types, starting from simple syntactic structures, followed by more complicated cases and including fully idiomatic expressions. Automatic evaluation shows that the described approach increases the quality of translation by 0.6 BLEU points.

1. Introduction

There are many cases in real texts when the meaning of collocation is not based on the meaning of its parts. Usually such phrases are called Multiword expressions (MWEs). MWEs include a large range of linguistic phenomena, such as nominal compounds, phrasal verbs, idiomatic expressions, terminology and institutionalized phrases.

Treatment of MWEs is one of the most complicated issues in natural language processing, especially in Machine Translation (MT). MWEs cannot be treated by general, compositional methods of linguistic analysis due to unclear semantics. Such approach causes overgeneration in cases when the meaning could be inferred from the words, e.g., 'telephone box' (Sag et al, 2002). Sag points to the idiomaticity problem for MWEs with opaque semantics: how to predict cases when MWE has a meaning which is unrelated to the meanings of its constituents (words), e.g., the meaning of idiom 'raining cats and dogs' is not related to 'cats' and 'dogs'.

Although meaning of MWEs cannot be derived from its component words, MWEs behave like any other phrase in a sentence, e.g., they take inflections, undergo syntactic operations etc.; at the same time, when MWE is translated, its syntactic structure in the translated phrase can be completely different from the source phrase.

Different strategies have been used for encoding of MWEs in different lexical resources. For languages with minimal inflection a lot of MWEs can be fixed in the lexicon as words with spaces. This approach is inappropriate for highly inflected languages with rather free word order where each MWE can have a lot of different morphological variants and can be used in the sentence in different syntactic roles.

Alvey Tools Lexicon (Carroll and Grover, 1989) provides good coverage of phrasal verbs with detailed information about syntactic aspects, but without distinguishing compositional from non-compositional entries and not specifying entries that can be productively formed. WordNet (Fellbaum, 1998) covers a large number of MWEs, but does not provide information about their variability. Neither of these resources covers idioms (Villavicencio et al., 2004).

According to Villavicencio, "the challenge in designing adequate lexical resources for MWEs, is to ensure that the variability and the extra dimensions required by the different types of MWE can be captured". Calzolari et al. (2002) focus on MWEs that are productive and present regularities which can be generalised and applied to other classes of words with similar properties.

Following this approach, the paper proposes flexible architecture for a lexical encoding of MWEs, which allows the unified treatment of different kinds of MWE in the translation process, taking into account syntactic similarities. The described MWE dictionary is used in a commercial English-Latvian MT system (Skadiņš et al. 2007). Processing of MWEs is one of the modules in the system which allows identifying, translating and generating MWEs as part of the sentence.

2. The grammatical system of Latvian

Latvian belongs to the class of inflected languages which are the most complex from the point of view of morphology.

Latvian nouns are divided into 6 declensions. Nouns and pronouns have 6 cases in both singular and plural. Adjectives, numerals and participles have 6 cases in singular and plural, 2 genders, and the definite and indefinite form. The rules of case generation differ for each group.

There are two numbers, three persons and three tenses (present, future and past tenses), both simple and compound, and 5 moods in the Latvian conjugation system.

Latvian is quite regular in the sense of forming inflected forms however the form endings in Latvian are highly ambiguous. Nouns in Latvian have 29 graphically different endings and only 13 of them are unambiguous, adjectives have 24 graphically different endings and half

of them are ambiguous, verbs have 28 graphically different endings and only 17 of them are unambiguous.
Another significant feature of Latvian is the relatively free word order in the sentence which makes parsing and translation complicated.
Like other languages, Latvian has a large number of MWEs. There are different types of MWEs in English which have to be translated into Latvian, e.g., phrasal verbs (e.g. “give up” – “padoties”, “slow down” – “piebremzēt”), nominal compounds (e.g. “telephone box” – “telefona būdiņa”), institutionalized phrases (e.g. “salt and pepper”) or phrases with truly idiomatic meaning (e.g. “early bird gets the worm” – “kurš putniņš agri ceļas, agri slauka deguntiņu”). There are cases when a single word in English should be translated in Latvian as a MWE (e.g. "arson" - "ļaunprātīga dedzināšana") and vice versa (e.g. "send word" – "paziņot").

3. MWE dictionary

Transfer of source language syntactic structures into the corresponding target language syntactic structures during the translation process could be implemented in many different ways. Mel'čuks lexical functions (LFs) (Mel’čuk, 1974) establish a semantic relation between one word or word combination, which is called function argument, and another word or word combination, which is called function value corresponding to this argument. LFs are universal regarding the language and therefore the translation could be acquired by identifying the arguments and the value of the LF during parsing and by substituting with the correct value from the target language dictionary during generation (Apresjan et al).
A different approach is the usage of Lexicalized Tree Adjoining Grammar (LTAG) (Abeillé et al, 1990). The transfer between two languages can be realized by directly putting large elementary units into correspondence without going through interlingual representation and without major changes to the source and target grammars. Transfer rules are stated as correspondences between nodes of trees which are associated with words.
Our dictionary of MWEs consists of a lexicon of phrases and a set of MWE rules. The lexical entry consists of a normalized source language MWE, its translation equivalent and an identifier of MWE rule describing syntactic structures of the source and the target MWE (see Table 1). Usually one rule describes tens, hundreds or even thousands of MWEs. Depending on the syntactic structure of the MWE, normalized MWE could be a list of the words in a base form or/and inflected or conjugated forms of the words.
The MWE rule describes the syntactic structure of MWE in the source language and its transformation into the corresponding structure of the target language.

Source phrase	Target phrase	Rule ID
talk around	runāt apkārt	V-ADV-7
clever boots	slīpēts zellis	A-N-9
have a swim	izpeldēties	V-DET-N-1
get a cold	saaukstēties	V-DET-N-1
sound a false note	uzņemt nepareizu toni	V-DET-A-N-14
out of temper	saniknots	ADV-PREP-N-1
have lunch	ēst pusdienas	d-V-N-3

Table 1: Lexicon of phrases

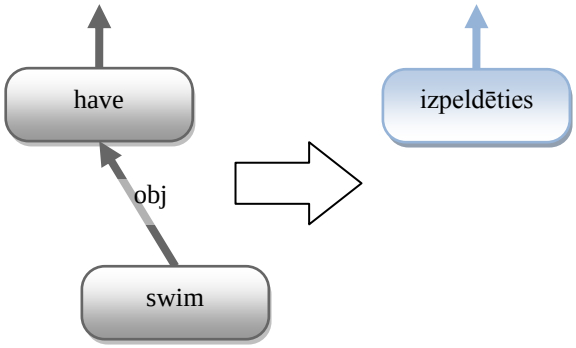
In simplest cases the source and target MWEs have the same syntactic structure and translations of words are attached to the corresponding nodes of syntactic tree. Example (1) shows the rule for such type of MWEs consisting of a main verb (V) and an adverb (ADV). It starts with the rule identifier V-ADV-7 followed by the syntactic structure of MWE in the source and target language (V1[advl:ADV2]=>V1[advl:ADV2]) and providing characteristics of the normalized phrase, e.g., V1.SourceBaseform stands for the verb in base form, ADV2.SourceSpelling stands for the adverb in its written form.

```
(1)   IdiomRule(V-ADV-7)
V1[advl:ADV2]=> V1[advl:ADV2]
{
      V1.SourceBaseform;
      ADV2.SourceSpelling;

      V1.TargetBaseform;
      ADV2.TargetSpelling;
}
```
talk(V1) around(ADV2) => runāt(V1) aplinkus(ADV2)

The MWE rule can also include morphological restrictions for a certain source language parse tree node and assign morphological features for a certain target language parse tree node. Example (2) shows the rule for the English noun phrase ’clever boots’ in plural and the corresponding Latvian noun phrase ‘slīpēts zellis’ in singular.

```
(2)   IdiomRule(A-N-9)
N2[attr:A1]=> N3[attr:A1]
{
      N2.Number == plural;
      A1.SourceSpelling;
      N2.SourceSpelling;

      A1.TargetBaseform;
      N3.TargetBaseform;
      N3.Number = singular;
}
```
clever(A1) boots(N2) => slīpēts(A1) zellis(N3)

Although the simplest MWEs form a considerable part of the MWE dictionary, most of the MWE rules are more complicated and describe transformation of parse tree between the source and target languages. Some nodes can be dropped from the source tree, some new ones can be added in the target tree during a transfer. In the most complicated cases the head node of the fragment tree can be changed into a different one. Figure 1 shows how English MWE 'have a swim' is transformed into a single Latvian word 'izpeldēties'.



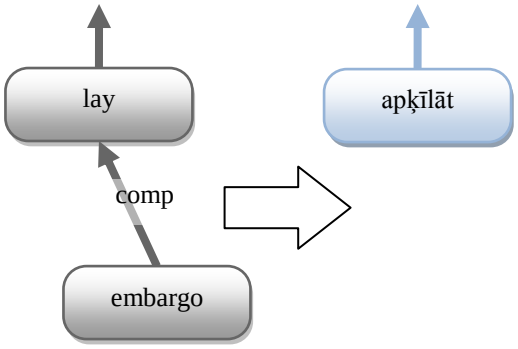
IdiomRule(V-DET-N-1)
V1[obj:N2[?det:DET3,null]]=>V1
{
V1.SourceBaseform;
DET3.SourceSpelling;
N2.SourceBaseform;

V1.TargetBaseform;
}

have(v1) swim(N2) => izpeldēties(V1)

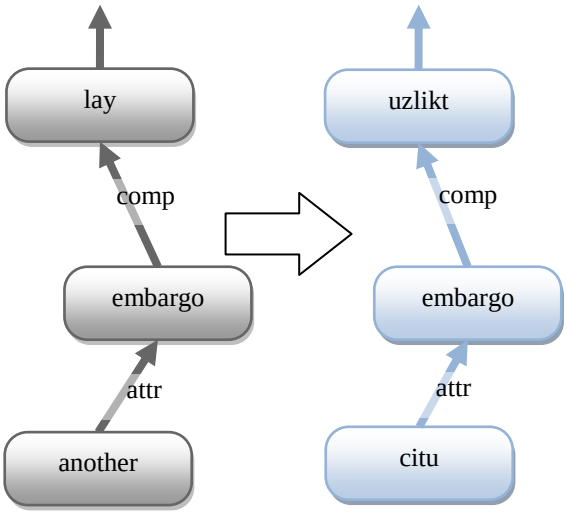
Figure 1: MWE rule where the source and target tree have different syntactic structures

Similar syntactic structures can be translated differently depending on the context they are used. Figure 2.1 and Figure 2.2 shows the translation process for the MWE 'lay an embargo' in two cases: as a single verb 'apķīlāt' or a verb phrase 'uzlikt embargo'. The rule from the Figure 2.1 will be applied only if the noun node N2 has no other children as the only optional determiner DET3; in this case the translation is a single verb and we can drop the N2 node in the target tree. In other cases the rule from the Figure 2.2 is performed, i.e., the same tree structure is kept.



V1[comp:N2[?det:DET3,null]]=>V1

Figure 2.1: Translation of 'to lay an embargo': similar syntactic structures in source language have different target language tree

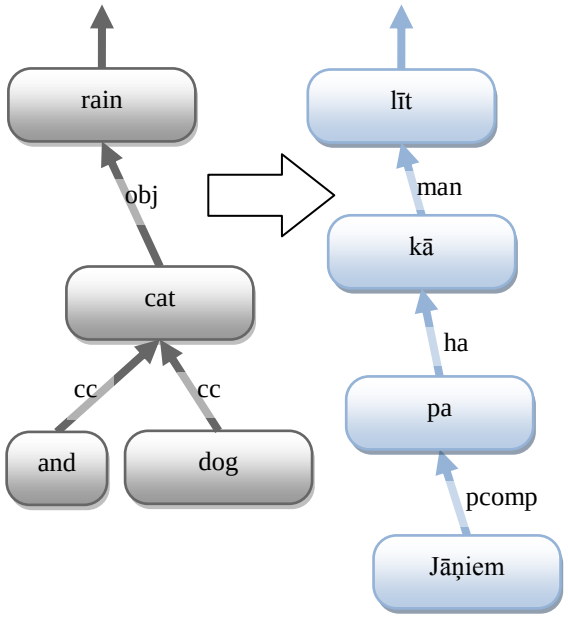


V1[comp:N2[?det:DET3]]=>V1[comp(right):N2]

Figure 2.2: Translation of 'to lay an embargo': similar syntactic structures in source language have different target language tree

Not only the structure of the syntactic tree, but also the word order can be changed during the translation process. Therefore, in the description of the target language tree, we specify not only the parse tree and the syntactic relations but also the word order, i.e., the position of the child node in respect to the parent node. The child node can be inserted directly before the parent ('left'), at the beginning of phrase ('leftmost'), directly after the parent ('right') or at the end of phrase ('rightmost').

Truly idiomatic expressions have completely different phrase structure in the source and the target languages. Figure 3 illustrates the translation of the idiom 'raining cats and dogs' into 'līst kā pa Jāņiem' ('it's raining like on Midsummer's Day'). Only the main verb node V1 is kept in target tree during the transfer, all other target nodes hasvebeen replaced with different ones.



V1[obj:N2[cc:CC3,cc:N4]] =>
V1[man(right):PART5[ha(right):PREP6[pcomp(right):N7]]]

Figure 3: Source and target tree for idiomatic expressions

4. Processing of Multiword Expressions in MT system

The English-Latvian MT system is built from separate components, each of them having their own functionality. Components are executed successively during the translation process: the system detects the language of the source text, builds the syntactic parse tree, performs MWE processing, performs syntactic and lexical transfer, disambiguates word translations, and establishes morphological agreement between words.

Input of the MWE module is the parse tree of the source language sentence. The MWE processing module traverses the parse tree top-down trying to identify the potential MWEs, i.e., patterns (fragments of parse tree) defined in MWE rules. If a match is found, the MWE rule looks for a lexical match in the lexicon of phrases. If the matching entry is found in the lexicon of phrases, the target tree fragment is created and lexical translations are attached to the right nodes.

The translated MWE is integrated into the target tree to be used later during transfer, agreement and other processes. In these modules MWE is treated in the same way as other words in sentence (conjugated, declined, etc.) to create a fluent target language sentence.

5. Results and Evaluation

The current MWE dictionary has a lexicon of 19,790 English MWEs with their translations, and 914 rules. The most frequent phrases are adjective-noun phrases (6995 entries), noun-noun phrases (3912 entries), verb-noun phrases (2597 entries), noun-preposition-noun phrases (1674), and verb-preposition-verb phrases (1010 entries).

We have compiled a English-Latvian corpus consisting of original sentences (500 units) for MT evaluation, i.e., this corpus is natural and sufficient for evaluation purposes. The compiled corpus is parallel (sentence-aligned), not annotated (morphologically, syntactically, and lexically unmarked), and is representational and balanced at the same time.

Two popular evaluation metrics NIST (Doddington, 2002) and BLEU (Papineni et al., 2002) were chosen for automatic evaluation. The evaluation results for MWE processing module in English-Latvian MT are summarized in Table 2.

System characteristics	BLEU	NIST
With MWE processing	18.17	4.7543
Without MWE processing	17.52	4.6802

Table 2: Evaluation results for English-Latvian MT

BLEU score rose by 0.6 points while NIST score rose by 0.07 points when MWE processing module was included. MWE processing module detected and provided translations for 83 MWEs in the test corpus.

6. Future work

We have developed an advanced MWE processing technique. It is included in the system and the evaluation shows improvement of the translation quality, but we see that the quality can be significantly improved by adding new MWEs. At the moment we use manually created rules and dictionaries for the MWE processing which is time consuming and expensive. We are planning to use unsupervised machine learning techniques to learn MWE rules and create dictionaries from the parallel corpus.

7. References

Abeille, A., Schabes, Y., Joshi, A. (1990). Using Lexicalized Tags forMachine Translation. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING 90),* Helsinki, Finland, pp. 1-6.

Apresjan, J.D., Boguslavsky, I.M., Iomdin, L.L., Tsinman, L.L. (2002). Lexical Functions in NLP: Possible Uses, Computational Linguistics for the New Millennium: Divergence or Synergy? In Manfred Klenner / Henriëtte Visser (eds.) *Proceedings of the International Symposium held at the Ruprecht-Karls-Universität Heidelberg, 21-22 July 2000*, Frankfurt am Main, pp. 55-72.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Canary Islands.

Carroll, J., Claire, G.. (1989). The derivation of a large computational lexicon of English from. LDOCE. In B. Boguraev and E. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing*. Longman.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of the Human Language Technology Conference (HLT)*, San Diego, CA, pp. 128—132.

Fellbaum, C. (1998). Towards a representation of idioms in WordNet. In *Proceedings of the workshop on the use of WordNet in Natural Language Processing Systems (Coling-ACL 1998)*, Montreal.

Mel'čuk, I. (1974). Opyt teorii lingvisticheskix modelej "Smysl ↔ Tekst", Nauka, Moscow.

Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword Expressions: a pain in the neck for NLP. In: *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City, Mexico, pp. 1–15.

Skadiņš, R., Skadiņa, I., Deksne, D., Gornostay, T. (2007). English/Russian-Latvian Machine Translation System. In: *Proceeding of the Third Baltic Conference on Human Language Technologies*. In press.

Villavicencio, A., Copestake, A., Waldron, B., Lambeau, F. (2004). Lexical Encoding of MWEs. In T.Tanaka, A. Villavicencio, F. Bond, A. Korhonen (Eds.) *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing. Barcelona.*