Analysis of Multiword Expression translation errors in Statistical Machine Translation

Natalia Klyueva
Institute of Formal and Applied Linguistics
Charles University in Prague
kljueva@ufal.mff.cuni.cz

Jeevanthi Liyanapathirana
Copenhagen Business School
Denmark
jlibc@cbs.dk

1. Introduction

Motivation SMT systems make errors in MWE, we search the ways how to improve it.

Noun multiword expressions : **En/Fr: military coup ||| coup d'etat**

Auxiliary multiword expressions: **En/Fr: with regard to ||| en ce qui concerne**

Light verbs: **Cs/Ru: dát smysl (give sense) ||| иметь смысл (have sense)**

Idioms: **En/Fr: kick the bucket ||| casser sa pipe (literal meaning)**

En/Fr: kick the bucket ||| mourir (figurative meaning)

Multiword Expressions (MWEs) present a sequence of words with non-compositional meaning, they differ from language to language and are highly idiosyncratic. Even for the related languages we can not be sure if the structure of MWE is similar or not to say nothing about typologically different languages. In this paper, we are going to evaluate a statistical machine translation (SMT) system, Moses, trained for several language pairs to explore how it cope with multiword expression translation. We will experiment with **Czech-Russian**, **English-French** language pairs to make sure that our conclusions are as language-independent as possible.

MT systems make mistakes in idioms.

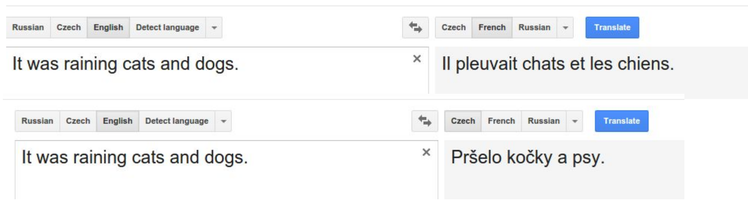
Moses:

entrées ||| entries (instead of ***ticket sales***)

Czech-Russian Moses:

návrh zákona (a bill) -> *работ закона ('work of projects') (correct is *законопроект* - 'lawproject')

... and Google Translate:



Moses: Czech-Russian

We integrate a list of MWE as additional data to the Czech-Russian language pair. We have checked the output of the baseline system, and there were quite a few mistakes in multiword expressions, especially in named entities.

A. Czech-Russian MWE from Wikipedia headlines

We used a list of names and phrases from Wikipedia headlines as this was the only parallel Czech-Russian resource of NEs we managed to obtain. The headlines were automatically extracted from the wikipedia dumps in XML (https://dumps.wikimedia.org/). The headlines were not necessarily multiword expressions, but for the sake of our experiment, we extracted MWEs.

Drawback: the data are not very clean and there are no idioms or light verbs.

MWE from wiki headlines

Inocenc IV. Иннокентий IV
Zub Зубы человека
Účtová osnova План счетов
Olaf III. Norský Олав III Тихий
Universal Mobile Telecommunications System UMTS
Lubrikační gel Лубрикант
Istrijská župa Истрийская жупания
Koliformní bakterie Колиморфные бактерии
Reigen Император Рэйгэн
Delta Dunaje Дельта Дуная
Gothic rock Готик-рок
Sfântu Gheorghe Сфынту-Георге
Projekt 949 Granit Подводные лодки проекта 949A «Антей»
Duševní vlastnictví Интеллектуальная собственность
Tel Aviv Тель-Авив
Messierův katalog Каталог Мессье
Generic Universal Role-Playing System GURPS
Fyzikální chemie Физическая химия
Turks a Caicos Тёркс и Кайкос
Zubní kartáček Зубная щётка
Koprivnicko-križevecká župa Копривницко-Крижевацкая жупания
Die Happy Die Happy
Dějiny Říma История Рима
Higašijama Император Хигасияма
Štika obecná Щука
Vánoční stromek Новогодняя ёлка
Křižák obecný Крестовик обыкновенный
Bosenskohercegovská hymna Гимн Боснии и Герцеговины
Gaius Licinius Macer Гай Лициний Макр
Ryzec pravý Рыжик настоящий
Politický systém Francie Политическая структура Франции
Mealyho automat Автомат Мили
Kočka bažinná Камышовый кот
Švýcarská hymna Гимн Швейцарии
Zápach z úst Галитоз
Leon V. Arménský Лев V Армянин
Hubbleova klasifikace galaxií Последовательность Хаббла
Dopravna Раздельный пункт
Krevní plazma Плазма крови
Severní Evropa Северная Европа
Ankan Император Анкан
Dimmu Borgir Dimmu Borgir
Houses of the Molé Houses of the Molé
Filth Pig Filth Pig
Joshua Abraham Norton Нортон I
Tatra K2 Tatra K2
Registrované partnerství Гражданское партнёрство
Teorie grup Теория групп
Seznam plemen koček Список пород кошек

B. Adding MWEs as a parallel corpus

Using the factored configuration of Moses, we ran two experiments:

- the baseline with models trained on data without the Wikipedia headlines
- model trained on data including the headlines

Actually, it was the winning setup (A) for pair French-English - adding data as a parallel corpus.

Total number of MWE pairs: **87,354**

C. Performance

In addition to BLEU, we calculated the number of out-of-vocabulary (unknown, OOV) words - searching for non-Latin characters. In the first experiment, the BLEU score was 17.23% with 1216 OOV words. The BLEU score in the second experiment was slightly better - 17.90% with 1011 OOV words.

	BLEU	OOV
baseline	17,23%	1216
with mwe	17,90%	1011

D. Examples of phrases where MWE were improved

We examined the list of OOV words in the output from the two experiments. Among those 205 words/MWEs that were recognized in the second experiment, there were MWEs from the added resource, such as Carlo Ancelotti, Amschel Rothschild, alt soprán etc. The following MWEs were not translated or mistranslated in baseline, but were translated correctly according to the added data in the improved setup:

Volební právo	Активное избирательное право
Průkaz totožnosti	Идентификационные карты
Higgsův boson	Бозон Хиггса
Velký hadronový urychlovač	Большой адронный коллайдер
Paliativní péče	Паллиативная помощь
Praní špinavých peněz	Отмывание денег

Moses: French - English

We experiment Multi Word Expression translation in French to English .

We check the possibility of integrating the multi word expression translation into Moses decoder .

A. Extracting MWE pairs

Extract some multi word expressions by defining a set of rules to extract multi word expressions.

Use part of speech taggers for French and English to get the lemmas for the source and target text . Lemmas reduce data sparseness which exist when only words are used.

e.g. Adj-Noun : Plenary meeting / Libre circulation
Noun-Adj : Parlement europeen
Noun-Noun : Member state / Etat membre

We also add named entities, and idioms to the MWE list manually.
e.g. Middle East , In particular

B. Aligning target text to MWE s

GIZA++ alignment to get the alignment in the target side.

We use the higher frequency translation candidates in the corpus as the MWE candidates.

This way, we extract the MWE expressions in both source and target sides

C. Incorporating MWE knowledge in Moses

Three Methods :

A) We use the extracted MWE pairs as a parallel corpus in addition to the normal corpus, and re train the model

B) We hard code the parallel MWE s in the phrase table with a lexical probability of 1

C) We include a new feature into the Moses decoder (second pass : used in MERT training) . The feature just says whether the phrase has a MWE or not.

Performance

Dataset : Europarl Corpus (French to English) Tools Used : GIZA++ , Moses , TreeTagger , Stanford Parser

Baseline BLEU Score : 21.67
Method A : 21.88
Method B : 21.68
Method C : 19.2

This project was supported by the grant LM2010013.