

MT and MWEs

MWEs are a challenge for MT :

• Example 1 : Idiom

Source EN	my job was a piece of cake
Google TR	mon travail était un morceau de gâteau

• Example 2 : Phrasal verbs

Source EN	[...] turned the stereo on
Google-TR	[...] tourna la stéréo sur

→ Word-for-word translation .

Extraction of a specific corpus for evaluating MT of MWEs

▶ Create a List of MWEs based on electronic dictionaries available on the Net, Forums

▶ Using LIDILEM's tool - Lexico-gramme [O. Kraif et S. Diwersy (2012)] . Europarl, News and Emolex (literary texts) corpora are indexed

Lexicogramme		Graphiques			
Show 25 entries					
I1	I2	f	f1	f2	am.log.likelihood
cut_*	off_ADV	2193	37692	20717	20719,9900
cut_*	speaker_NOUN	985	37692	19153	7802,7565
cut_*	president_NOUN	995	37692	57057	5723,1800
cut_*	back_ADV	478	37692	43937	2301,1423
cut_*	down_ADV	409	37692	43867	1846,0920

▶ Corpus specific extracted from Emolex "Phrasal Verbs" with word from 0 à 5 .

▶ Select sentences that pose problems to Google TR at the MWE level

▶ 500 sentences pairs (Tst-MWE) :

- 40 different phrasal verbs (366 sentences)
- 74 different idioms (134 sentences)

▶ Extract a control corpus of 500 random sentences (Tst-Normal)

Pre-processing of PV for En-Fr SMT training

▶ Identification of PVs of our test corpus using the parts of speech and some dependencies provided by XIP [Ait-Mokhtar et al., 2002]

▶ Gluing verb and prep for training and test (Tst-MWE)

▶ Examples :

keep your voice **down**, Hermione begged him .

keep-down your voice, Hermione begged him .

He picked up a remote and **turned** the stereo **on**

He picked up a remote and **turned-on** the stereo

▶ Problem : This pre-processing is not straightforward (need for MWE detection)

Results

	Corpus of 500 sent.		Detail per corpus or per category on Tst-MWE				Idioms only in Tst-MWE #134 sent
	Tst-NORMAL	Tst-MWE	PV only #366 sent	Litt. Texts #272 sent	Europarl #213 sent	News #15 sent	
Moses-LIG	24.87%	20.83%	22.72%	12.59%	29.33%	21.94%	15.21%
+pre-proc	23.81%	21.19%	23.18%	14.10%	29.38%	19.80%	14.79%
Google-TR	19,27%	19,81%	18.67%	10,9%	21,82%	12,00%	19.75%

BLEU scores (Moses-LIG vs Google)

- Slight but not significant improvement
- Need an evaluation metric that could "zoom" on MWEs
- BLEU [Papineni and al , 2002] is calculated on the whole corpus → do not see the effect on smaller events

Processing of idioms

▶ Using a pre-defined « dictionary » of idioms

▶ Constrained SMT decoding [koehn,2014]

▶ Example :

And you thought my job was a **piece of cake** !

And you thought my job was a
<idiom translation="facile"> **piece of cake** </idiom> !

→ BLEU goes from 15.21 % to 30.71 % on the 134 sentences containing idioms

→ We used a small pre-defined list of idioms that covered 100 % of our test set → not very realistic

→ Idiom detection is not addressed here (pre-defined list)

Conclusion

▶ Corpora with bigger density of MWEs than in usual text

▶ Need for an automatic metric that focuses on specific events such as MWEs (otherwise we do not see much differences)

▶ First propositions to handle MWE

→ Pre-processing of PV for En-Fr SMT

→ Use constrained decoding for Idioms

▶ Example :

Source	surely they must call the operation off now ?
Reference	maintenant, ils doivent sûrement annuler l'opération .
Hyp (baseline)	ils doivent appeler l'opération maintenant ?
Hyp (+pre-proc)	ils doivent annuler le fonctionnement maintenant ?

Abstract

Multiword Expressions (MWE) are complex phenomena which represent a real challenge for computational linguistics, and Statistical Machine Translation systems (SMT) in particular. In this paper, we try to improve translation of Multiword Expressions in SMT. Preliminary results, regarding the BLEU score, are encouraging.

References

- [Kraif and Diwersy, 2012]: "Le Lexicoscope: un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques" p 399406, TALN June 2012, Grenoble France.
- [Besacier et al., 2012]: "The LIG English to French Machine Translation System for IWSLT 2012", In proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT).
- [Papineni and al , 2002]: "BLEU: a Method for Automatic Evaluation of Machine Translation", July 2002, pages 311-318, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia.
- [Koehn, 2014], "Statistical Machine Translation, System User Manual and Code Guide", University of Edinburgh Royaume-Uni .