Identifying Multi-Word Expressions from Parallel Corpora with String-Kernels and Alignment Methods

WG1 - WG3

Parseme 5th GM
Iași, 23-24 September 2015

Johanna Monti
jmonti@uniss.it

Federico Sangati
federico.sangati@gmail.com

Mihael Arcan
mihael.arcan@deri.org

Non-Translatability

Identification of MWE based on **non-literal translatability property**
an MWE cannot be translated from one language to another on a word-for-word basis.

Property of MWEs with limited or no variation of distribution, such as:

- idioms (*it's raining cats and dogs* → it. **sta piovendo cani e gatti*)
- many collocations (*heavy rain* → it. **pioggia pesante*)
- fixed expressions (*by and large* → it. **da e largo*)
- proverbs (*there's no place like home* → it. **non c'è posto come casa*)
- phrasal verbs (*Bring somebody down* → it. **Portare qualcuno giù*)

Translating MWEs implies several problems due to their morpho-syntactic, semantic and pragmatic idiomaticity.

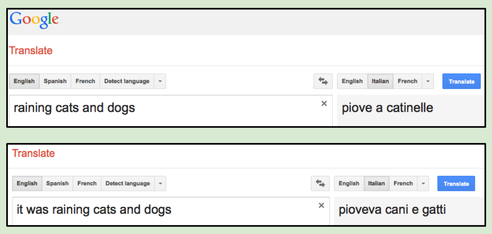
MWEs are sometimes **discontinuous**, i.e. it is possible to insert an element between the constituents of a MWE , e.g. *it's looking really good*.

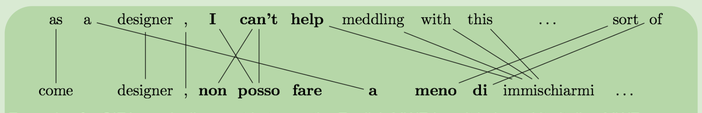
Translational asymmetries: differences between an MWE in the source language and its translation in the target language.

We can have different correspondences between languages:

- many-to-many (en. *kick the bucket* → it. *tirare le cuoia*)
- many-to-one (en. *kick the bucket* → it. *morire*)
- one-to-many (it. *svegliare*→ en. *wake up*)

Limitation of current PBSMT systems





Example of a GIZA++ misalignment between the English MWE I can't help and its Italian MWE translation non posso fare a meno di (lit. not can do to less than). Dashed lines are those not connecting the MWEs in the source and target sentence.

MANUALLY

Annotating the English-Italian TED parallel corpus (**WIT³**, see Cettolo et al., 2012) with MWEs. We used the 2010 test subset (1529 sentences).

3 PHASES

1. Individual annotation: 13 annotators, each sentence assigned to at least 2 annotators.

2. Inter-annotation check: each annotator confirms or changes the annotations after being confronted with the others' annotations.

3. Validation: integration and resolution of possible annotation conflicts.

AUTOMATICALLY

Two-stage process:

1. Identifying a list of *potential MWEs* pairs via **Parallel String-Kernel** (including discontinuous sequences).
2. Filter out those candidates which are *not MWEs* using
   a. alignment information
   b. co-occurrence statistics

Dataset:

- For training the SMT system we use the 2014 released **WIT³ TED data set**, which contains ~190K parallel sentences (for training).
- We using on the 2014 TED **development** set (~1K sentences) and the 2010/2011/2012 **test** sets (~1.5K sentences, each).

1. Individual Annotation

SNT #	Source (EN)	MANUAL Manual Translation (IT)	AUTO Automatic Translation (IT)	MWE SOURCE TEXT	MWE MANUAL TEXT	MWE MANUAL CHECK (Y/N)	MWE AUTO TEXT	MWE AUTO CHECK (Y/N)
369	people sort of think i went away between " titanic " and " avatar " and was buffing my nails someplace , sitting at the beach .	la gente pensa quasi che me ne sia andato tra " titanic " e " avatar " e che mi stessi girando i pollici seduto su qualche spiaggia .	persone come pensare partii tra " titanic " e " avatar " e fu buffing mie unghie da qualche parte , seduto in spiaggia .	buffing my nails	girando i pollici	Y	buffing mie unghie	N

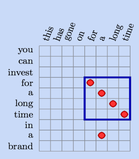
3. Validation

SNT #	Source (EN)	MANUAL Manual Translation (IT)	AUTO Automatic Translation (IT)	ANN #	MWE SOURCE TEXT	MWE MANUAL TEXT	MWE MANUAL CHECK (Y/N)	MWE AUTO TEXT	MWE AUTO CHECK (Y/N)
26	" don , " i said , " " don " , gli ho detto " tanto per capire bene , voi siete famosi per fare allevamento così lontano , in mare aperto , che non inquinate . "	" non " , ho detto " , per ottenere i fatti dritto , siete famosa per coltivare così lontano in mare , non inquinante . "						
				3	to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N
				9	just to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N
				13	get...stright	capire bene	Y	per ottenere...dritto	N
				FINAL	just to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N

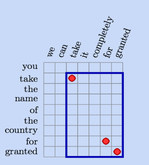
Methodology

PHASE 1 (identification)

English	Italian
Make her feel special, don't take the relationship for granted.	Falla sentire speciale, non dare il rapporto per scontato.
You take results for granted because you have seen it happen so often.	Sbagliato dare i risultati per scontato.



Contiguous MWE: **for a long time**



Discontiguous MWE: **take [...] for granted**

PHASE 2 (Filtering)

a. Alignment Step

The filter checks that both conditions are met:

1. The two expressions are not perfectly aligned (to avoid word-for-word aligned expressions).
2. There is at least a word in A and B aligned (to avoid case of complete misalignment which are typical signs of accidental matches)

b. Statistical Step

Use of the frequency counts of the extracted pairs to compute conditional probabilities:

$$\alpha = P(B|A) = \frac{freq(A,B)}{freq(A)}$$

$$\beta = P(A|B) = \frac{freq(A,B)}{freq(B)}$$

where A and B are source and target sequences.

Annotation EVALUATION

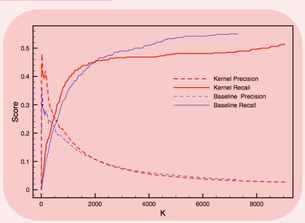
- Annotation Agreement:
  - At least two annotators agreed for the 27% (671) of the MWEs and in 45% of them (1,115) at least two annotators showed an overlapping (at least one word in common).
- Final annotation:
  - 799 English MWE types (931 tokens), of which 729 (91%) are contiguous and the 9% (70) are discontinuous.
  - Most MWEs have length of 2 (515) and 3 (261), but there are MWEs up to the length of 8.
  - In 52% of the cases (471) the annotators have evaluated the automatic translation to be incorrect.

English	Italian
pointed at	indicò
no longer	non ... più
don 't get me wrong	non fraintendetemi
got bitten by	sono stato affetto dal
a lot of	un sacco di
in the dead of winter	nella tristezza dell' inverno

Identification/MT EVALUATION

1. MWE Identification:

Our approach reaches higher level of precision and recall with respect to a standard baseline extraction methods (Pointwise Mutual Information).



2. Machine Translation:

The best results are obtained if we use the extracted bilingual MWEs with a threshold of 0.8. The average improvement of all three test sets is more than 0.4 BLEU points.

	Test Set 2010				Test Set 2011				Test Set 2012			
Threshold (σ)	0.0	0.2	0.4	0.8	0.0	0.2	0.4	0.8	0.0	0.2	0.4	0.8
Multi TM (f)	23.52	23.76	23.91	**24.31***	23.64	23.68	23.77	23.70	24.06	24.16	24.10	24.22
CB ageing (f)	22.58	24.06	23.82	23.29	22.75	23.90	23.65	**24.05***	23.35	24.50*	24.10	**24.55***
Multi TM	22.36	23.44	23.73	24.02	22.27	23.25	23.66	23.94*	22.92	23.71	24.09	24.43*
CB ageing	22.35	23.03	22.75	24.22	22.62	22.94	22.67	23.33	23.19	23.66	23.35	23.64
Baseline	23.97				23.61				23.97			