FONDAZIONE
BRUNO KESSLER

MWE: I can't help falling in love with you

Federico Sangati
FBK-DH, Trento
sangati@fbk.eu

Andreas van Cranenburgh
Huygens ING, Royal Netherlands Academy of Arts & Sciences; ILLC, Univ. of Amsterdam.
andreas.van.cranenburgh@huygens.knaw.nl

Johanna Monti
Sassari University
jmonti@uniss.it

Institute for Logic, Language and Computation
University of Amsterdam

uniss
UNIVERSITÀ DEGLI STUDI DI SASSARI

Abstract

We investigate new ways for identifying MWEs from parallel multi-lingual corpora, based on the **non-translatability** property of MWEs: an MWE cannot be translated from one language to another on a word by word basis (Sag et al., 2002; Monti, 2012).

Phase 1: Kernel Methods

Goal: Identify potential MWEs in parallel pairs of sentences (in one language, in the other, or in both).

Input: large bilingual corpus sentence aligned.

Kernel methodology:

- For every pair of sentences in the corpus, the algorithm will detect a pair of sentences in the source language which **share** a certain expression, for which the correspondent pair of sentences in the target language also **share** an expression.
- Can be computed efficiently via **string kernel** (Lodhi et al., 2002) for aligned text, while **tree kernels** can be employed (Sangati et al., 2010; van Cranenburgh, 2014) if a parallel treebank is available.

Example:

English	Italian
I feel we will have to **call it a day** at this point.	Credo che a questo punto dobbiamo **passare oltre**.
He would like us to adjourn the vote to the next part-session and **call it a day** for now.	Il relatore chiede di rinviare la votazione alla prossima seduta e, per ora, di **passare oltre**.

Outcome cases:

	English	Italian
1.	**MWE** *bring up to date*	× *modernizzare*
2.	× *he died*	**MWE** *ha tirato le cuoia*
3.	**MWE** *call it a day*	**MWE** *passare oltre*
4.	× *aims at adapting*	× *mira ad adattare*

PARSEME Working Groups

WG1: Lexicon-Grammar Interface Depelopment of linguistic resources, MWE dictionaries.

WG3: Statistical, Hybrid and Multilingual Processing of MWEs Hybrid methodology for MWE identification and translation.

MWE and Non-Translatability

Target cases

Fixed expressions e.g., EN. by and large → IT. *da e largo.

Idioms e.g., EN. Call it a day → IT. *Chiamarlo un giorno.

Proverbs e.g., EN. There's no such thing as a free lunch → IT. *Non esiste una cosa come un pranzo gratuito.

Phrasal verbs e.g., EN. Bring somebody down → IT. *Portare qualcuno giù.

Exceptions

A number of MWEs can be translated literally to all other languages, such as proper names and universal proverbs. These are therefore excluded from the scope of the current work.

Phase 2: MT Filtering

Goal: remove candidate pairs without MWEs.

- Phase 1 is prone to find many pairs of candidate expressions which do not include MWEs (e.g., last row of outcome cases).

Methods:

- Traditional "word by word" translation system (detect which candidate pairs are literal translations).
- 1:1 alignement pairs between source and target languages obtained via GIZA++ (Och and Ney, 2003).

Phase 3: Crowdsourcing

Goal: validate the final list of candidate pairs using crowdsourcing methods:

- Amazon Mechanical Turk
- CrowdFlower
- Educ. tools for second language learners
- CAT systems for human translators

Related Work

Some recent approaches rely on the exploitation of the translational correspondences of MWEs.

De Medeiros Caseli et al. (2010) identification of MWEs in a multilingual context, exploiting a word alignment process. Also associates some multiword expressions with semantics.

Tsvetkov and Wintner (2014) exploit non-compositional translation of MWEs and developed a new alignment-based algorithm for MWE extraction focused on misalignments, augmented by validating statistics computed from a monolingual corpus.

Segura and Prince (2014) propose an alignment process between pairs of sentences, strongly based on syntax. It relies on a rule-based system combining partial alignments from a database through a non-iterative graph-theory based process.

Arcan et al. (2014) address the problems of automatic identification of bilingual terminology using Wikipedia as a lexical resource, and its integration into an SMT system using the XML mark-up and the Fill-Up model methods.

Case Study: can't help

Corpus: TED Talks EN-IT (Cettolo et al., 2012)

- Number of sentences: 187,809
- Tokenized and aligned with GIZA++ (many thanks to Mihael Arcan)

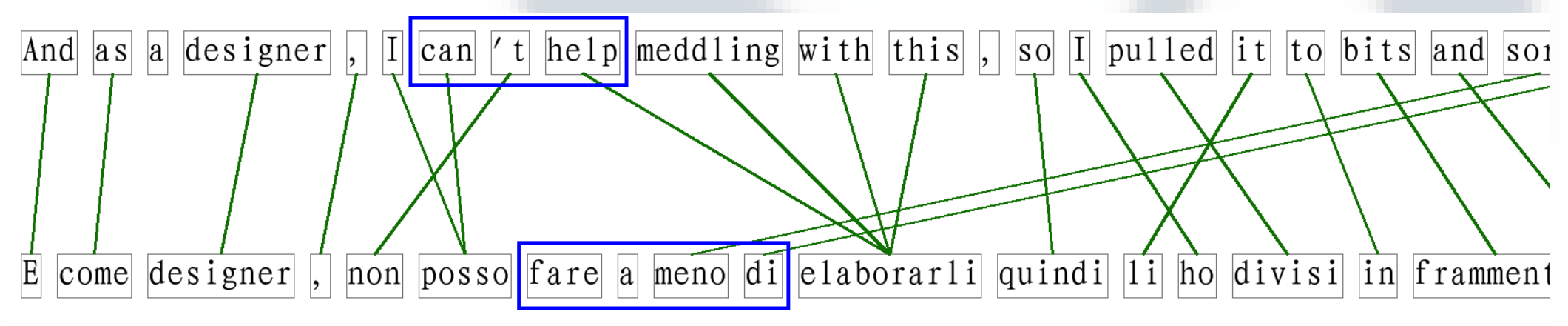
Target MWE: EN: **can't help** → IT: **fare a meno di**

Corpus Analysis:

- Intersection (4) [1760, 41845, 87214, 107792]
- Only in EN (22) [9303, 9316, 13677, 13687, 15336, 22592, ...]
- Only in IT (7) [41031, 41213, 46509, 101575, 117009, 161383, 165466]

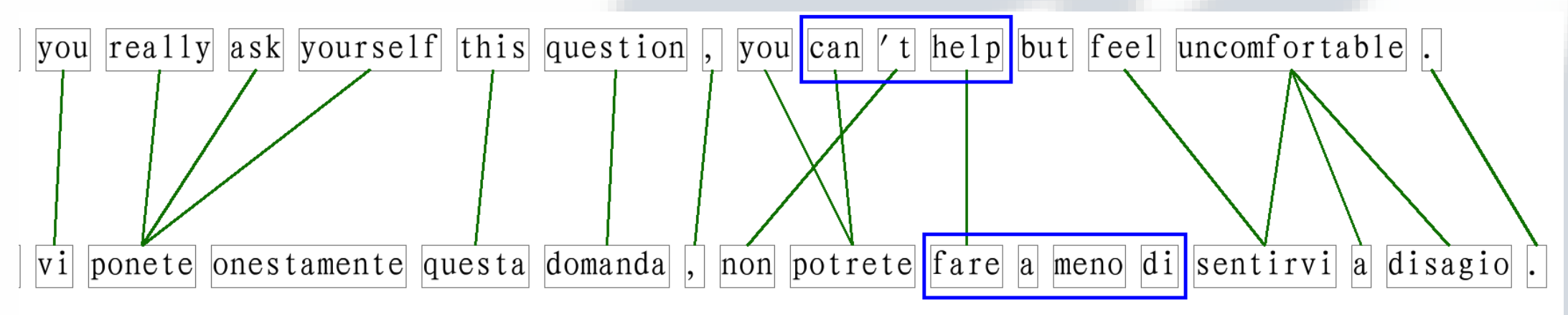
GIZA++ alignements:

1760



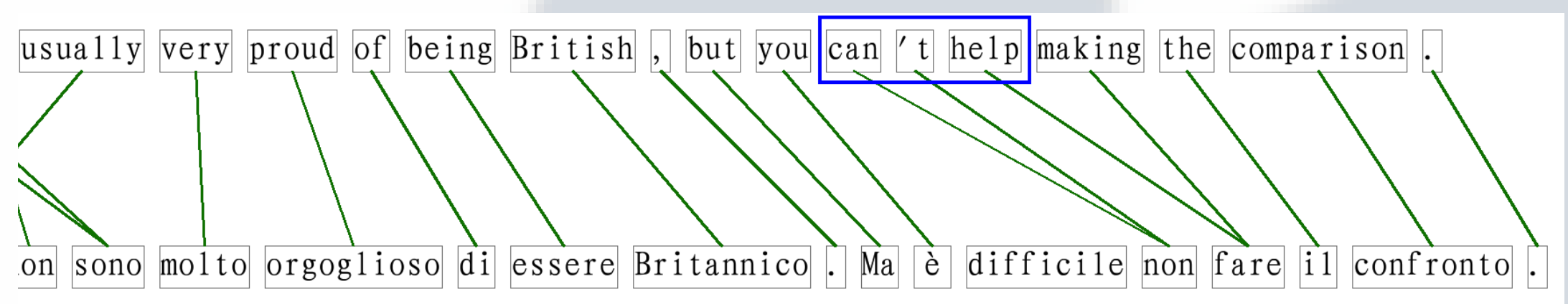
And as a designer , I can 't help meddling with this , so I pulled it to bits and so
E come designer , non posso fare a meno di elaborarli quindi li ho divisi in framment

107792



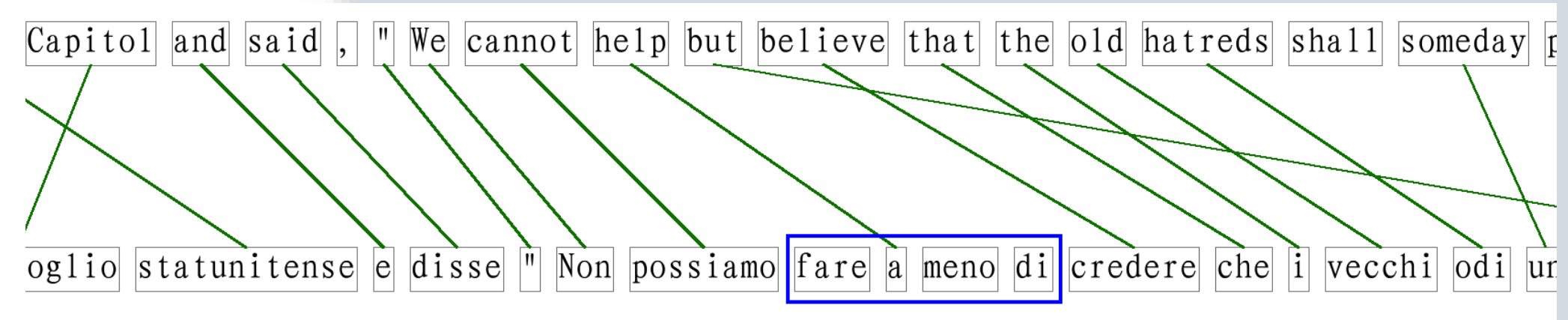
you really ask yourself this question , you can 't help but feel uncomfortable .
vi ponete onestamente questa domanda , non potrete fare a meno di sentirvi a disagio .

9303



usually very proud of being British , but you can 't help making the comparison .
on sono molto orgoglioso di essere Britannico . Ma è difficile non fare il confronto .

41031



Capitol and said , " We cannot help but believe that the old hatreds shall someday p
oglio statunitense e disse " Non possiamo fare a meno di credere che i vecchi odi ur

MT Systems:
EN source: "I **can't help** falling in love with you."

Google (2014.09.01)	Correct
* Non posso **fare a innamorarsi** di te.	Non posso **fare a meno di** innamorarmi di te.

References

Mihael Arcan, Claudio Giuliano, Marco Turchi, and Paul Buitelaar. 2014. Identification of bilingual terms from monolingual documents for statistical machine translation. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268. Trento, Italy.

Helena de Medeiros Caseli, Carlos Ramisch, Maria das Gracas Volpe Nune, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2):59–77. URL http://opus.bath.ac.uk/18664/.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419—444.

Johanna Monti. 2012. *Multi-word unit processing in Machine Translation - Developing and using language resources for Multi-word unit processing in Machine Translation*. Ph.D. thesis, University of Salerno.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg. URL http://dx.doi.org/10.1007/3-540-45715-1_1.

Federico Sangati, Willem Zuidema, and Rens Bod. 2010. Efficiently Extract Recurring Tree Fragments from Large Treebanks. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

Johan Segura and Violaine Prince. 2014. Using Alignment to detect associated multiword expressions in bilingual corpora. Tralogy [En ligne], Tralogy I, Session 6 - Translation and Natural Language Processing / Traduction et traitement automatique des langues (TAL).

Yulia Tsvetkov and Shuly Wintner. 2014. Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics*, 40(2):449–468.

Andreas van Cranenburgh. 2014. Linear average time extraction of phrase-structure fragments. *Computational Linguistics in the Netherlands Journal*, x(accepted for publication):x–y.