The Name of the Game is Comparable Corpora

Ruslan Mitkov
Research Group in Computational Linguistics
University of Wolverhampton
R.Mitkov@wlv.ac.uk

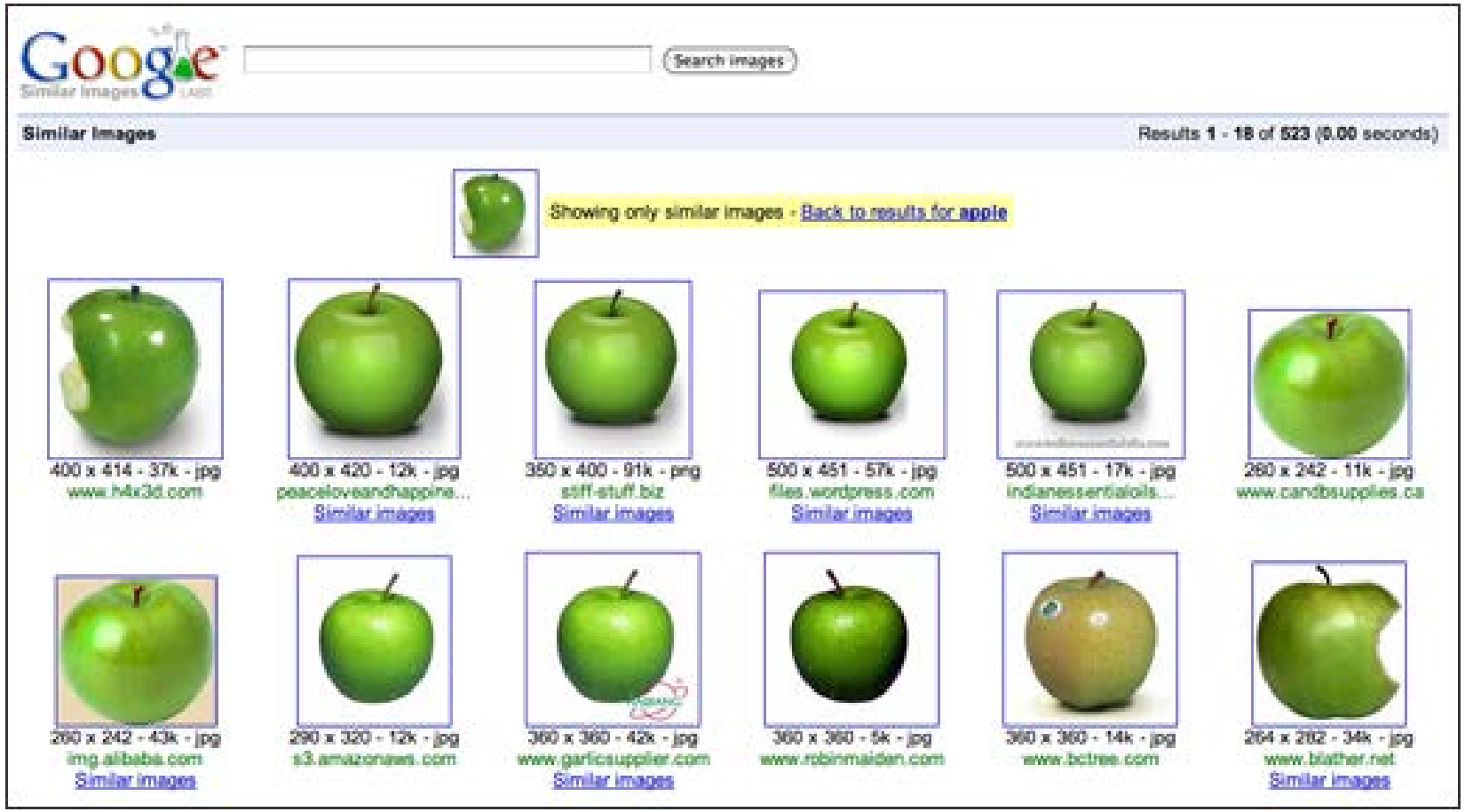
RGCL
Research Group in Computational Linguistics

Preliminaries

Comparable corpora: when are corpora ‘comparable’?

Basic concepts and definitions

When are corpora ‘comparable’?



RGCL
Research Group in Computational Linguistics

Comparable corpora (McEnery, 2003)

- *Comparable corpora* are corpora where a series of monolingual corpora are collected for a range of languages, preferably using the same sampling frame and with similar balance and representativeness, to enable the study of those languages in contrast.

Sampling frame

- The sampling frame is essential (Tony McEnery *2006)*:
- The components representing the languages involved must match each other in terms of proportion, genre, domain and sampling period

Degree of comparability

Degrees of non-parallel corpora (Fung 2004, 2005)

- parallel corpora - sentence-aligned corpus containing bilingual translations of the same document
- noisy parallel corpora - non-aligned sentences which are nevertheless mostly bilingual translations of the same document
- comparable corpora - non-sentence-aligned, non-translated bilingual documents which are topic-aligned.
- quasi-comparable corpora - disparate, very-non-parallel bilingual documents which could either be on the same topic (in-topic) or not (off-topic)

Multilingual applications and corpora

- Ideally, parallel data would be the best resource both for multilingual NLP applications and for users.
- However, parallel corpora or translation memories may not be available, difficult to acquire or may be time-consuming to develop.
- Alternative and more promising approach would be to benefit from comparable corpora.

Comparable corpora are…

- … the most versatile and valuable resource for multilingual Natural Language Processing
- … and ‘multilingual’ language users

Multilingual NLP applications

- Machine Translation (see Rapp, Sharoff and Zweigenbaum 2016)
- Word translation (Rapp 1995; Gaussier et al. 2004; Gamillo and Pichel 2007; Pekar, Mitkov et al. 2008)
- Term extraction (Fung and McKeown 1997; Daille and Morin 2005; Saralegi, San Vicente and Gurrutxaga 2008)
- Bilingual document similarity (Sharoff, Zweigenbaum and Rapp 2015; Jagarlamundi et al. 2010)
- Crosslingual coreference resolution (Green at al. 2011)
- Name entity transliteration (Udupa et al. 2008; Klementiev and Roth 2006)
- Other multilingual applications such as
  - Automatic identification of cognates and false friends (Mitkov et al. 2008)
  - Testing the validity of translation universals (Corpas, Mitkov et al. 2008)
  - Tracking language change (Stajner and Mitkov 2012; Stajner, Mitkov and Leech 2013)
  - Automatic extraction and translation of multiword expressions (Mitkov 2016; Taslimipoor, Mitkov, Corpas and Fazly 2016)

Language users

- Translators (Zanettin 1998; Saldahna and O'Brien 2002; Olohan 2002; Corpas and Seghiri 2009; Corpas and Seghiri 2016)
- Terminologists (Lemay et al. 2005; Durán Muñoz 2012)
- Interpreters (Pérez Pérez 2013)

Comparable corpora within and across languages

- Kilgariff (2003): comparable corpora may be of the same or different languages
- Regards ‘comparability’ as ‘similarity’

Food for thought

Is surface similarity the best way forward?

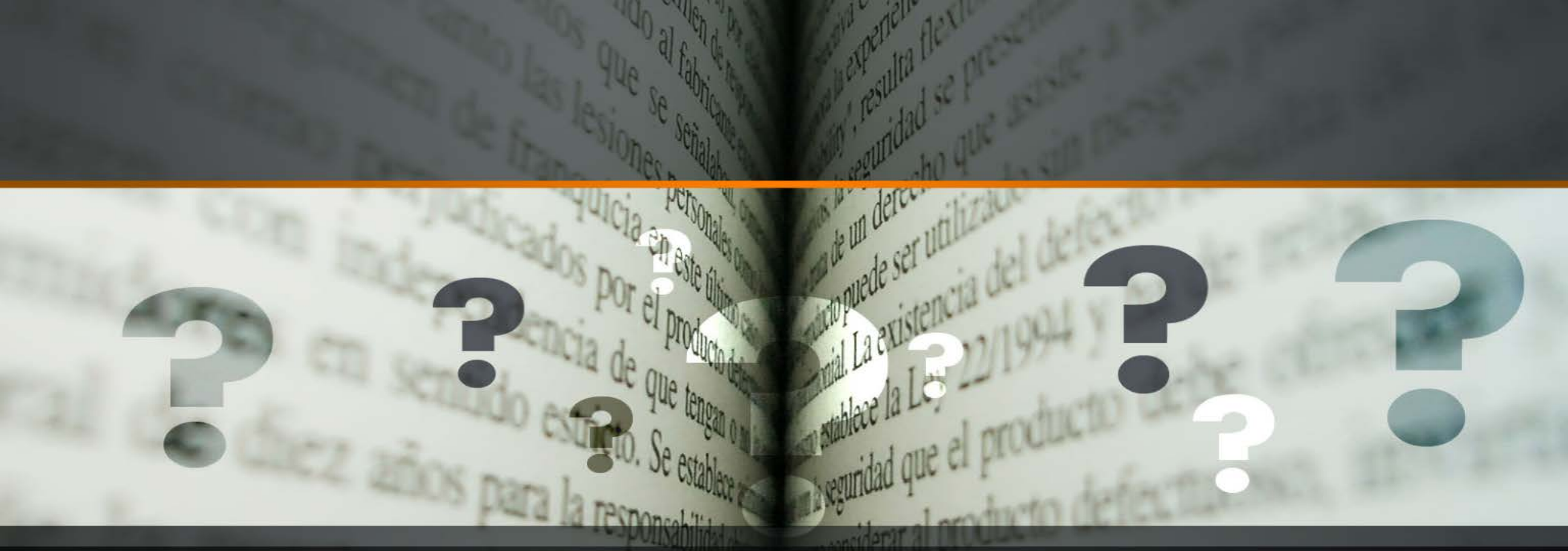
- Comparable corpora are usually compiled using surface similarity (statistical) techniques.
- Is this the best way forward?
- Example from the field of Translation Memory (character-string similarity, calculated through Levenshtein distance).

TM example

- SDL Trados gives the segments ‘Prendre des mesures de dotation et de classification.’ and ‘Connaissance des techniques de rédaction et de révision.’ a match rating of 56% because half of the words are the same and they are in the same position, even though the words in common are only function words (Gow 2003).

A better way forward

- Is semantic similarity (to include similarity of words, sentences, topics, documents …) a better way forward?
- However, is it feasible to compile comparable corpora on the basis of semantic similarity?



The new Revolution in the translation industry? Next generation Translation Memory systems

Ruslan Mitkov

I like Alicante which is such an attractive and exciting place.

I love Alicante as the city is full of attractions and excitements.

I dislike Alicante which is such an unattractive and unexciting place.

Sentence A	Sentence B	STS	Edit Distance
I like Alicante which is such an attractive and exciting place.	I love Alicante as the city is full of attractions and excitements.	3	72

Sentence A	Sentence B	STS	Edit Distance
I like Alicante which is such an attractive and exciting place	I dislike Alicante which is such an unattractive and unexciting place	1	92

Moving in the right direction…

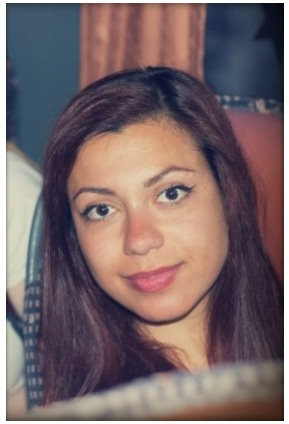
Sentence A	Sentence B	STS	Edit Distance
I like Alicante which is such an attractive and exciting place.	I love Alicante as the city is full of attractions and excitements.	3	72
I like Alicante which is such an attractive and exciting place	I dislike Alicante which is such an unattractive and unexciting place	1	92

The last word

- Comparable corpora are the most realistic, versatile and valuable resource for multilingual Natural Language Processing
- Comparable corpora are the safest and most promising resource for translators too
- Comparable corpora can offer more in terms of value and can support a wider range of applications and users
- The Name of the Game in Multilingual NLP is ‘Comparable Corpora’

Acknowledgements

Yvonne Skalban



Raya Petrova

RGCL
Research Group in Computational Linguistics

The Name of the Game is Comparable Corpora

Ruslan Mitkov
Research Group in Computational Linguistics
University of Wolverhampton
R.Mitkov@wlv.ac.uk

RGCL
Research Group in Computational Linguistics