Figurative Language in Big Data

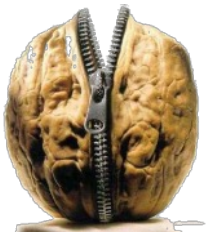
TraMOOC

Translation for Massive Open Online Courses

@IITPA, NLP Connect Talk Series

TraMOOC
Translation for Massive Open Online Courses

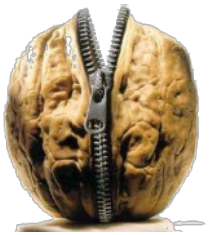
TraMOOC in a nutshell



Objectives & expected impacts

- TraMOOC has made existing monolingual educational material available to speakers of other languages.
- The project's vision has been to tear down language barriers, thus providing previously excluded groups of people with new educational chances.
- The project results has been showcased and tested on the openHPI platform and on the VideoLectures.Net digital video lecture library.
- The core of the service is open-source, with some premium add-on services which will be commercialised.
- The translation methodology is automatic and language-independent in nature and showcased for 11 indicative language pairs - 9 EU (DE, IT, PT, DU, BG, EL, PL, CS and CR), and 2 BRIC languages (RU and ZH).

TraMOOC in a nutshell

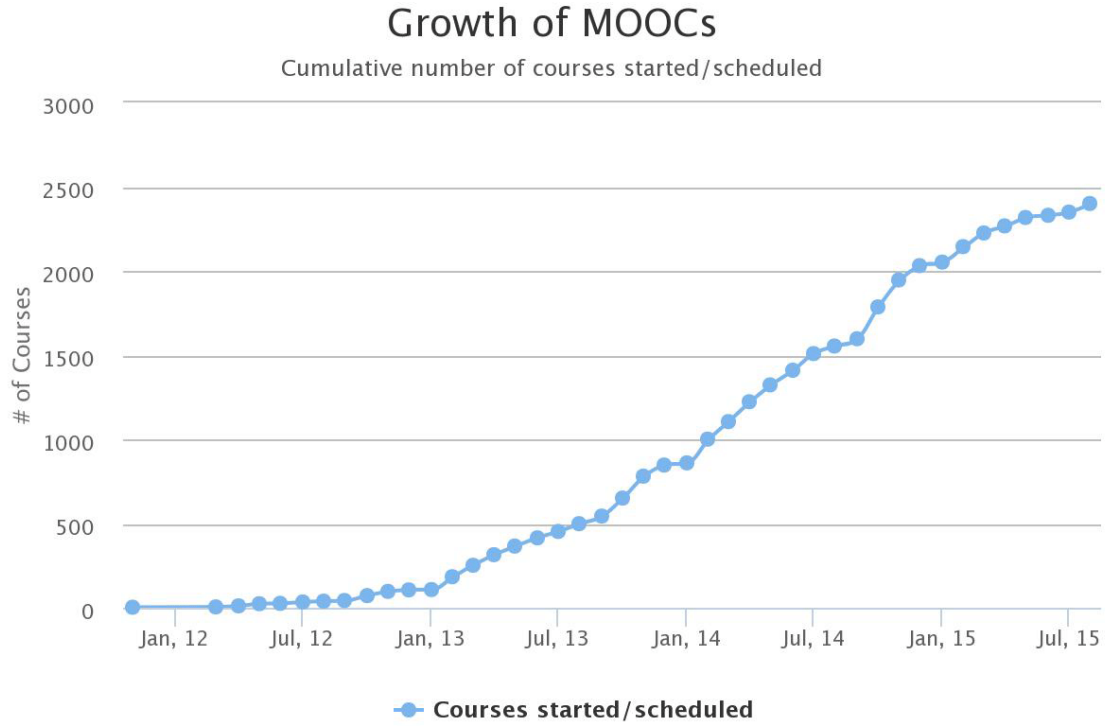


Main novelties

- Novel research in online and open education
  - Novel translation evaluation schemata
  - Added value to existing tools and resources in linguistics, natural language processing, text analytics, data mining and machine translation scientific communities
  - Topic identification of the source and translated text
  - Sentiment analysis on users' posts on fora and social websites has been used for extracting users' opinion on the translated material

Project Motivation

- MOOCs have been growing rapidly in size and impact



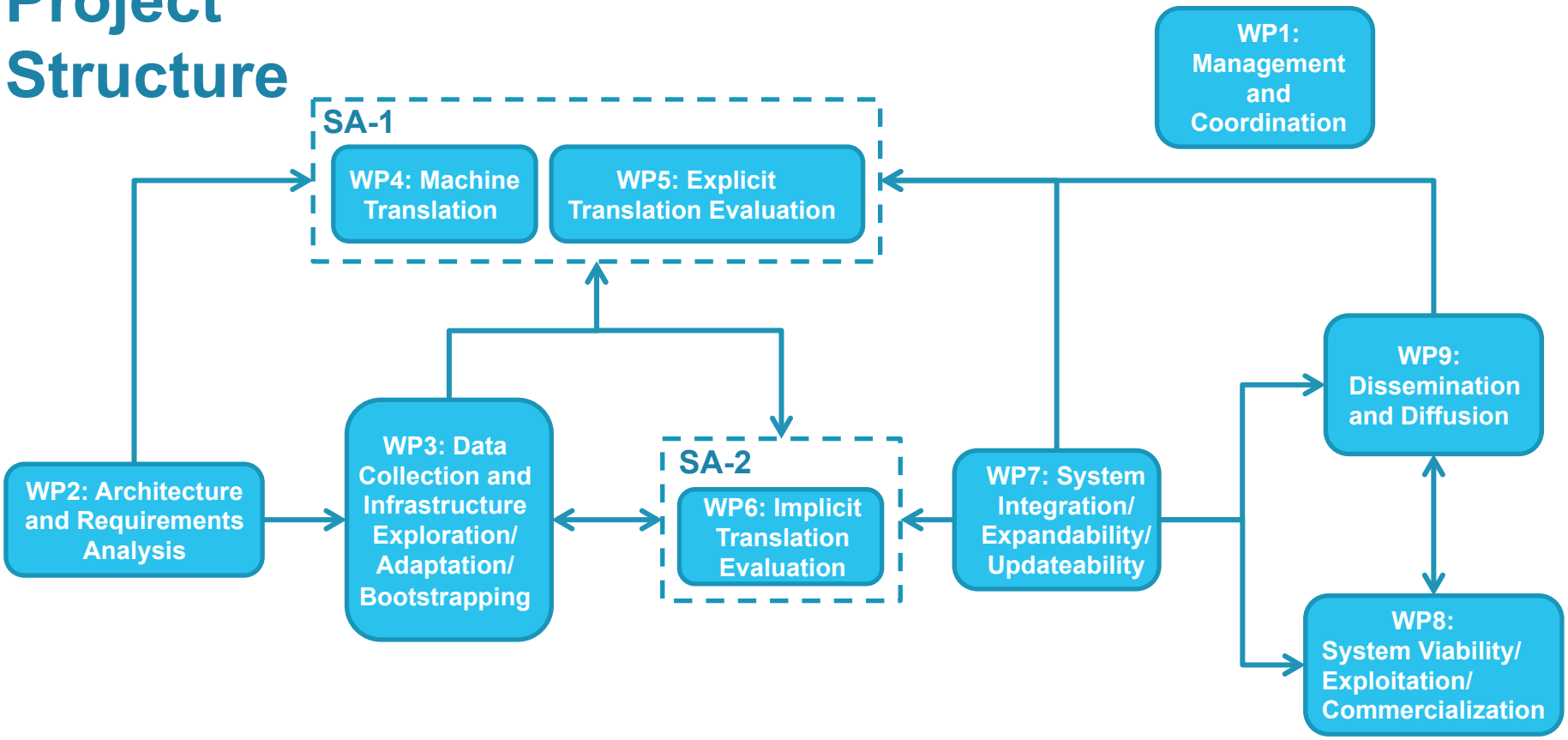
This year, the number of universities offering MOOCs has doubled to exceed 400 universities, with a doubling of the number of cumulative courses offered, to 2400. It is estimated that 16-18 million students attend MOOCs worldwide.*

* Source: https://www.edsurge.com/n/2014-12-26-moocs-in-2014-breaking-down-the-numbers

Work Description

HOW?

Project Structure



TraMOOC
Translation for Massive Open Online Courses

The TraMOOC Consortium

- TraMOOC brings together a consortium of leading researchers, highly relevant industry organizations and leading use-case partners.
- The partners' diverse interests in machine translation, linguistics, text mining web analytics and crowdsourcing methodologies-related areas make the consortium ideally placed to tackle the challenges associated with TraMOOC.
- The design of the scientific areas and the associated work packages have been arranged carefully to ensure maximum efficiency of input from each partner while maintaining a suitable distribution of responsibilities.

Crowdsourcing Platform Selection

Multiple platforms have been researched and ranked.

CrowdFlower was ranked second best after Amazon Mechanical Turk, the latter being rejected due to its inflexible USA-based payment process.

CrowdFlower was selected due to its:

- configurability,
- robust infrastructure,
- densely populated crowd channels and the evaluation and ranking process they undergo,
- convenient payment options,
- high reception and popularity level in the microtasking field

Crowdsourcing Platform Configuration

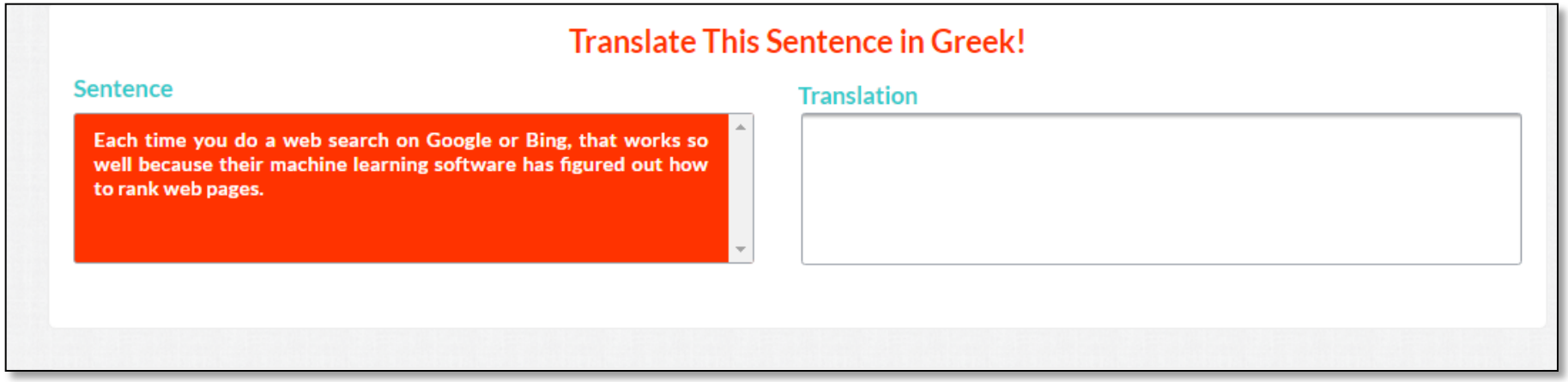
Crowdsourcing Activities:

1. CA1: Translation
2. CA2: Translation Evaluation
3. CA3: Sentiment/Topic Annotation

Crowdsourcing Platform Configuration
CA1: Translation

Activity: Parallel translation of EN segments to 11 target languages, 11M segments.

Data Sources: iversity, Coursera, QED Corpus



Data Processing

- Tokenization
- Sentence segmentation
- Truecasing
- Assure proper sentence alignment
- Marking of URLs
- Other steps specific to each data source (e.g., conversion from PDF into plain text)
- Goal: well-tokenized and as much as possible well-segmented grammatical parallel data
- Generally performed using a pipeline of available sentence splitters/aligners, data source tailored Python and shell scripts
- All data stored in a protected data repository provided by UBER

Tuning and Test Sets

- Need for a uniform in-domain test set throughout the project
- 80,000 words extracted from MOOC materials provided by IVE and DME
- Manual translation into Greek, Italian, and Portuguese performed by the respective partners
- 3 of the 4 language pairs in MT prototype 1 covered
- The translations for the remaining 8 languages were produced via crowdsourcing

Issues with Data Collection

- Challenging crawling: often complicated structure of the web resource; did not allow for large-scale automatic crawling
- Challenging data extraction and alignment: most materials in PDF, possible misalignments during conversion into plain text
- Representativeness: slides, notes, assignments are rarely translated
- Copyright issues

In-domain Data Summary

Language pair	Size (million words)
EN-DE	2.7
EN-BG	1.5
EN-PT	4.8
EN-EL	2.4
EN-NL	1.3
EN-CZ	1.5
EN-RU	1.4
EN-CR	0.2
EN-PL	1.7
EN-IT	2.3
EN-ZH	8.7

Bootstrapping of Data/Resources

- A case study with Croatian and Serbian
- Vanilla Moses trained on Coursera data in Croatian and Serbian shown to outperform a system trained only on Croatian in- and out-of-domain data
- High-quality MT of Serbian Coursera data into Croatian
- The Croatian data resulting from (rule-based) MT was added to the “normal” Croatian Coursera in-domain training corpus: best performing system

MWEs and Figurative Language

MWE and Metaphor Detection:

1. classify whether a multiword unit is metaphorical or not, given a target verb in a sentence; e.g., “The experts started *examining* the Soviet Union with a microscope to study perceived changes.”;
2. given a sentence, detect all of the metaphorical words (independent of their POS tags);
3. the aim has been to explore whether relatively standard architectures based on bi-directional LSTMs (Hochreiter and Schmidhuber, 1997) augmented with contextualized word embeddings (Peters et al., 2018) perform well on both tasks.

MWEs and Figurative Language

MWE and Metaphor Classification:

1. deep learning, and particularly CNNs, given the amount of data at our disposal;
2. the aim is also to verify the validity of the data intensive approach we have been advocating in the project;
3. the classification process should be independent of any syntactic patterns.

MWEs and Figurative Language

For MWE and Metaphor (Semantic) Analysis:

1. development of fine grained datasets, where metaphoricity is represented as a gradient property;
2. a classifier should have the ability to predict a degree of metaphoricity;
3. allow more fine-grained distinctions to be captured.

MWEs and Figurative Language

The nature of the Metaphor Interpretation Task (following Bizzoni and Lappin, 2018)

1. present a new kind of corpus to evaluate metaphor paraphrase detection;
2. construct a novel type of DNN architecture for a set of metaphor interpretation tasks;
3. Come up with a model that learns an effective representation of sentences, starting from the distributional representations of their words;
4. use word embeddings trained on very large corpora.

MWEs and Figurative Language

The nature of the Metaphor Interpretation Task (following Bizzoni and Lappin, 2018)

1. The model should be able to retrieve from the original semantic spaces not only the primary meaning or denotation of words, but also some of the more subtle semantic aspects involved in the metaphorical use of terms;
2. Corpus design based on the view that paraphrase ranking is a useful way to approach the metaphor interpretation problem.

MWEs and Figurative Language

The nature of the Metaphor Interpretation Task (following Bizzoni and Lappin, 2018)

1. The neural network architecture they propose encodes each sentence in a 10 dimensional vector representation, combining a CNN, an LSTM RNN, and two densely connected neural layers.
2. The two input representations are merged through concatenation and fed to a series of densely connected layers.
3. They show that such an architecture is able, to an extent, to learn metaphor-to-literal paraphrase.
4. The model learns to classify a sentence as a valid or invalid literal interpretation of a given metaphor, and it retains enough information to assign a gradient value to sets of sentences in a way that correlates with the source annotation of the crowd used.

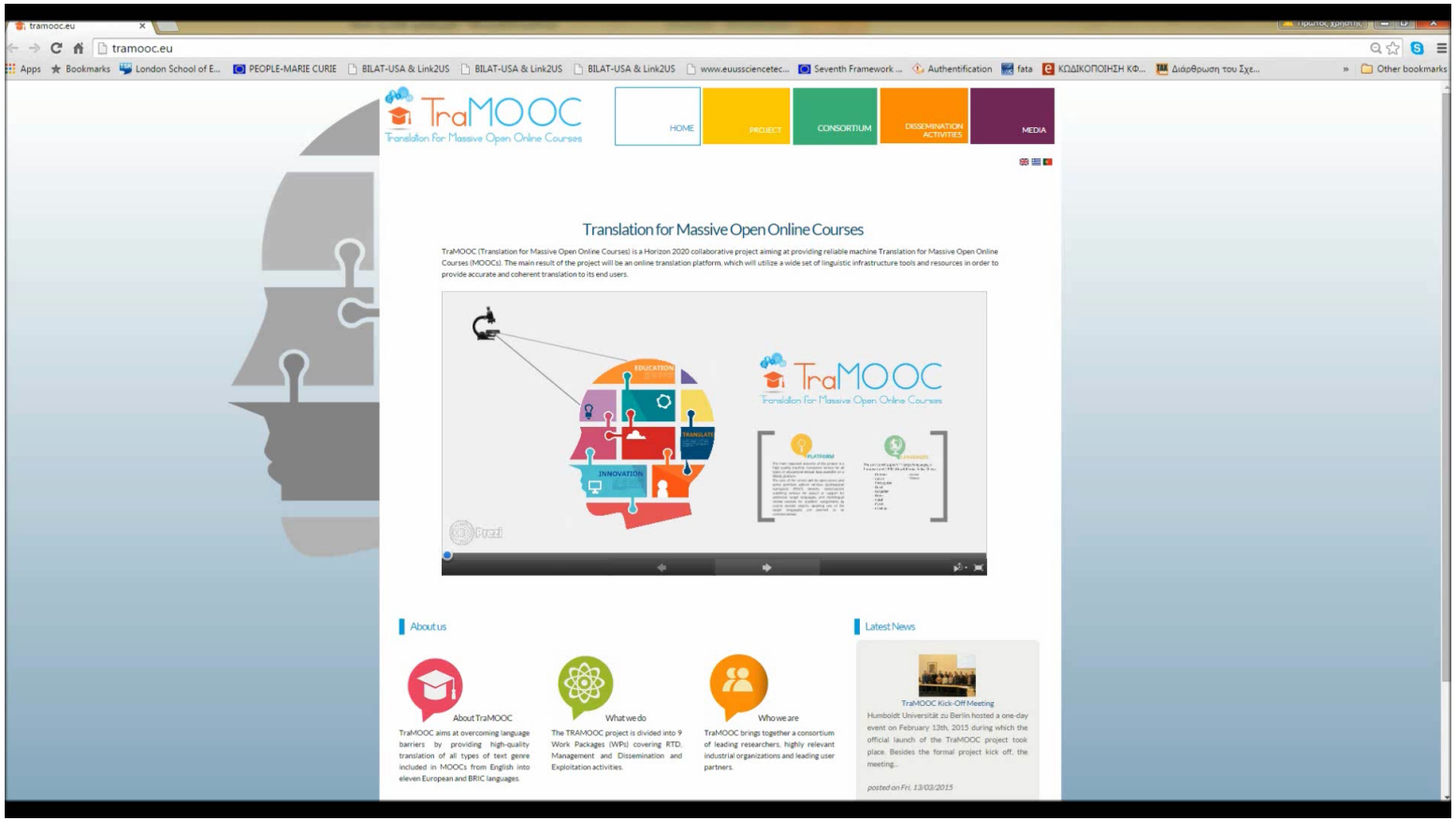
MWEs and Figurative Language

The nature of the Metaphor Interpretation Task (following Bizzoni and Lappin, 2018)

1. The model does not make use of any ”alignment” of the data.
2. The encoders’ representations are simply concatenated.
3. This gives the DNN considerable flexibility in modeling interpretation patterns.

The TraMOOC website

Find out more about the TraMOOC project and platform @ tramooc.eu



Valia Kordoni (UBER, TraMOOC Coordinator)