From MultiJEDI to MOUSSE: Two ERC Projects for Innovating Multilingual Disambiguation and Semantic Parsing of Text

Valerio Basile
Sapienza University of Rome
basile@di.uniroma1.it

Roberto Navigli
Sapienza University of Rome
navigli@di.uniroma1.it

ACM Reference Format:
Valerio Basile and Roberto Navigli. 2018. From MultiJEDI to MOUSSE: Two ERC Projects for Innovating Multilingual Disambiguation and Semantic Parsing of Text. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3184558.3186207

1 INTRODUCTION

The exponential growth of the Web is resulting in vast amounts of online content. However, the information expressed therein is not at easy reach: what we typically browse is only an infinitesimal part of the Web. And even if we had time to read all the Web we could not understand it, as most of it is written in languages we do not speak. Rather than time, a key problem for a machine is language comprehension, that is, enabling a machine to transform sentences, i.e., sequences of characters, into machine-readable semantic representations linked to existing meaning inventories such as computational lexicons and knowledge bases.

In this paper we present two interrelated projects funded by the European Research Council (ERC) aimed at addressing and overcoming the current limits of lexical semantics: MultiJEDI (Section 2) and MOUSSE (Section 4). We also present the results of Babelscape (Section 3), a Sapienza spin-off company with the goal of making the project outcomes sustainable in the long term.

2 ERC PROJECT: MULTIJEDI (2011-2016)

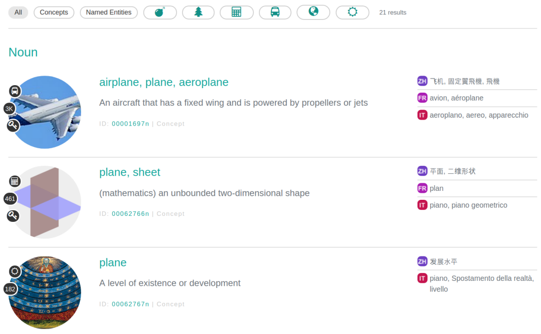
Focus. The *MultiJEDI* project[1] was funded by a 5-year ERC Starting Grant and coordinated by Prof. Roberto Navigli. It was focused on a long-standing problem in NLP, namely *word sense disambiguation (WSD)*, that is, the task of identifying the meaning of the words in a text, and linking them to a repository of lexical-semantic knowledge such as an electronic dictionary.

Limits. As of 2010, WSD was hampered by two key issues:

1. **Lack of supervision:** because disambiguation requires training word experts with hundreds of annotated sentences, performing the task at large scale requires millions of annotated sentences.
2. **Language coverage:** the main sense inventory available was WordNet for English, and coverage of other languages was limited, not to mention coverage of named entities.

2.1 Contribution 1: BabelNet

To overcome the above issues, as our first prominent contribution we devised a new resource which, rather than separating data by language, would synergistically interlink and integrate lexical-semantic information across languages. We therefore created *BabelNet*[2], a multilingual encyclopedic dictionary and semantic network which connects concepts and named entities in a very large network of semantic relations [2]. BabelNet has been initially created by merging two widely used resources, namely Wikipedia and WordNet, thus leveraging the multilinguality of the former and the semantic structure of the latter. Each Babel synset represents a given meaning and contains all the synonyms which express that meaning in a range of different languages. BabelNet is freely accessible through a Java API, an HTTP REST API, and a SPARQL endpoint[3]. Figure 1 shows an example of its Web interface.



Noun

airplane, plane, aeroplane
An aircraft that has a fixed wing and is powered by propellers or jets
ID: 00001697n | Concept
ZH 飞机, 固定翼飛機, 飛機
FR avion, aéroplane
IT aeroplano, aereo, apparecchio

plane, sheet
(mathematics) an unbounded two-dimensional shape
ID: 00062766n | Concept
ZH 平面, 二维形状
FR plan
IT piano, piano geometrico

plane
A level of existence or development
ID: 00062767n | Concept
ZH 发展水平
IT piano, Spostamento della realtà, livello

Figure 1: The "plane" entry in BabelNet with translations in 3 languages.

BabelNet is linked to widely used resources, including Wikipedia, Wiktionary, WordNet, and Omegawiki (see https://datahub.io/dataset/babelnet for the position of BabelNet in the Linked Open Data cloud), and it has been successfully employed for tasks such as computing multilingual semantic relatedness, word sense disambiguation and entity linking, and creating vectorial representations of concepts, among others.

Among its successes, BabelNet won the Prominent Paper Award 2017 from Artificial Intelligence, the most prestigious journal in the field of AI, the prestigious META prize 2015 for "groundbreaking work in overcoming language barriers", and was featured in an article in Time magazine.

[1] `http://multijedi.org/`

[2] `http://babelnet.org/`

[3] `http://babelnet.org/sparql`

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186207

Project	MultiJEDI	MOUSSE
Funding agency	European Research Council	European Research Council
Amount	1,288,400 EUR	1,497,250 EUR
Duration	From February 1st, 2011 to January 31st, 2016	From June 1st, 2017 to May 31st, 2022
Host institution	Sapienza University of Rome	Sapienza University of Rome
Team	Roberto Navigli (Principal Investigator), José Camacho Colados, Francesco Cecconi, Claudio Delli Bovi, Antonio Di Marco, Maud Ehrmann, Stefano Faralli, Tiziano Flati, Ignacio Iacobacci, David Jurgens, Andrea Moro, Mohammad Taher Pilehvar, Simone Ponzetto, Alessandro Raganato, Daniele Vannella	Roberto Navigli (P.I.), Valerio Basile, Andrea Di Fabio, Marco Maru, Valentina Piromalli, Tommaso Pasini, Valentina Pyatkin, Federico Scozzafava, more to come
Project website	`http://multijedi.org/`	`http://www.mousse-project.org/`

Table 1: The MultiJEDI and MOUSSE ERC projects at a glance.

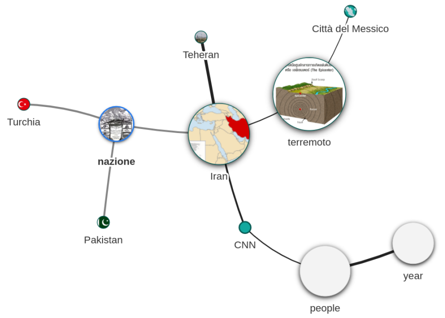


Figure 2: Concept and entity graph output by Extraggo from a news article on an earthquake in the Iran region.

2.2 Contribution 2: Multilingual WSD and EL

To overcome the paucity of semantically annotated data for disambiguation and enable multilinguality, we put forward a novel idea: leveraging our multilingual lexical knowledge resource, BabelNet, to perform state-of-the-art WSD and entity linking jointly. The resulting system, *Babelfy*[4] [1], works in any of the 271 languages supported by BabelNet.

3 INTERLUDE: BABELSCAPE

Towards the end of the project MultiJEDI, on the thrust of interested customers, we founded *Babelscape*[5], a Sapienza spin-off company created with the primary goal of making the group's research outcomes sustainable. Babelscape is now in full activity, with about 20 employees. Babelscape is currently working on:

- **BabelNet Live:** while BabelNet keeps being expanded and improved in terms of covered languages, quality of the mappings to other resources, and lexicalizations, in March 2017 the company launched *BabelNet Live*, a live edition which is continuously and automatically updated over time from its underlying resources (e.g., Wikipedia).
- **Extraggo:** Babelscape is also developing new tools on top of the MultiJEDI outcomes. Here we cite *Extraggo*[6], a system which extracts key concepts and entities from raw text, i.e. Babel synsets. Thanks to WordAtlas, a professional edition of BabelNet, the resulting semantic network can be viewed in any supported language (Figure 2 shows an example translated into Italian of a graph extracted from a CNN article).

4 ERC PROJECT: MOUSSE (2017-2022)

Focus. In June 2017 we started MOUSSE, a second 5-year ERC project, again coordinated by prof. Navigli, aimed to take multilingual text understanding to the next level. The groundbreaking goal of MOUSSE is to make a big leap forward and move from the analysis of natural language based on lexical units (i.e., words and multiword expressions, and their senses), to a formal representation of the meaning of sentences and larger texts.

Limits. This task is known as *semantic parsing*, and it is far from being a solved problem. Current semantic parsers require supervision, binding them to the language of interest and hindering their extension to multiple languages.

Contribution. The MOUSSE project will put forward innovative techniques to enable computers "to comprehend" texts in any language through the automatic creation of semantic phrase representations that are independent of the language used to express a given idea. The open issues in multilingual semantic parsing mirror the state of the art in WSD before MultiJEDI: there are few high-quality resources to build a semantic parser, focused mainly on English. With MOUSSE, we aim at bridging this gap and induce a paradigm shift that will enable full-fledged, language-independent computer understanding of natural language.

Since MOUSSE has recently started, no mature result is available yet. However, we are planning releases soon in 2018.

5 PRESENTATION AT WWW 2018

During the Web Conference 2018, a presentation will be given by Prof. Navigli on the results of the research efforts described here. The presentation will include a demo where all the systems described will be showcased; for the first time the preliminary results of the project MOUSSE will also be presented.

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487.

REFERENCES

[1] Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)* 2 (2014), 231–244.

[2] Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence* 193 (2012), 217–250.

[4] `http://babelfy.org`

[5] `http://babelscape.com`

[6] `http://extraggo.com`