A ColWordNet API

- Luis Espinosa-Anke
- Jose Camacho-Collados
- Sara Rodríguez-Fernández
- Horacio Saggion
- Leo Wanner

taln research group

UPF

SAPIENZA UNIVERSITÀ DI ROMA

A CoolWordNet API

- Luis Espinosa-Anke
- Jose Camacho-Collados
- Sara Rodríguez-Fernández
- Horacio Saggion
- Leo Wanner

taln research group

UPF

SAPIENZA UNIVERSITÀ DI ROMA

Outline

- Motivation
- ColWordNet
- The API

1. Motivation

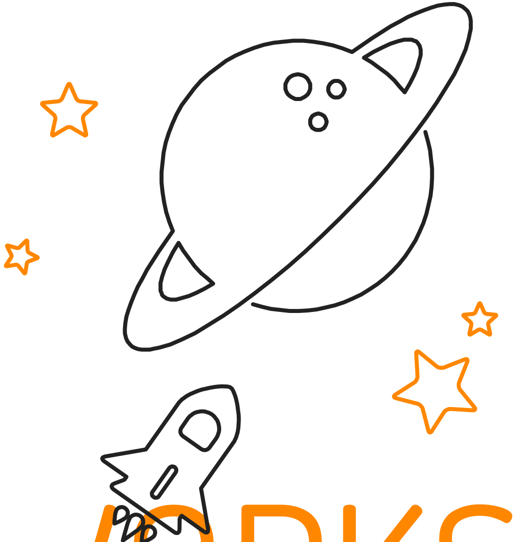
WordNet as a lexical resource for language learning and AI

Motivation: WordNet is a useful resource in NLP

- “The list of papers citing WordNet seems endless” (Hovy, Navigli and Ponzetto, AI 2013)
- It is a useful lexical resource for many tasks at different spectrums of LTs.
- One area where there is clear room for improvement is on *lexical combinations* of words: collocations.

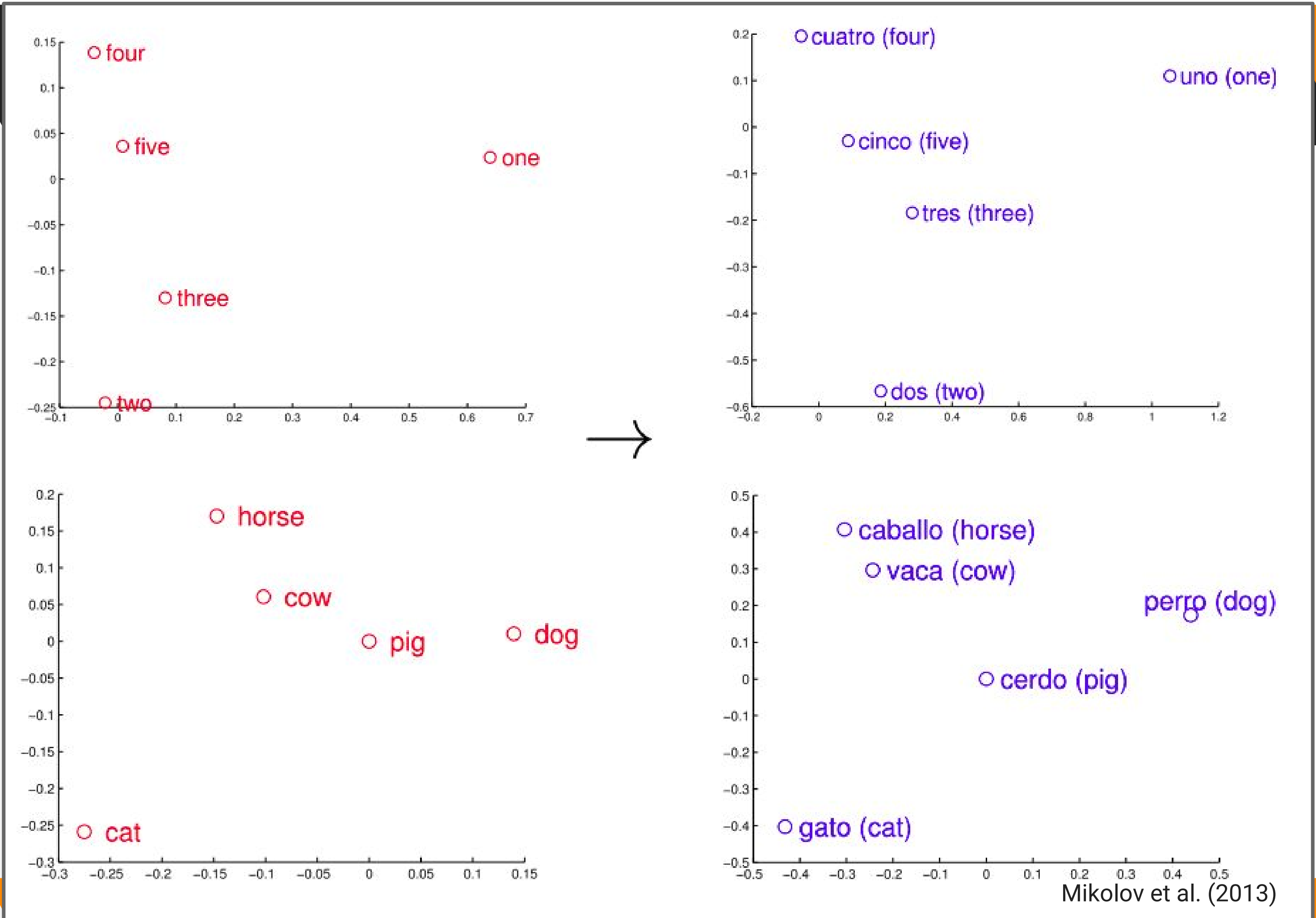
WordNet could be extended with collocational information

- In Espinosa-Anke et al. (Coling 2016) we describe and evaluate ColWordNet (CWN).
  - Previous work on collocation acquisition focuses on compiling collocation lists (Church and Hanks 1989, Kilgarriff 2006) - No semantic classification.
  - We tackled fine-grained collocation classification, drawing upon the lexical relation between the base and the collocate:
  - ‘**perform**’: take [an] exam, make [a] decision, pose [a] question
  - ‘**put an end**’: solve [a] problem, break the silence
  - Linguistic motivation based on the Meaning Text Theory (Melćuk, 1987)



CWN: HOW IT WORKS

A brief description of our method for creating CWN

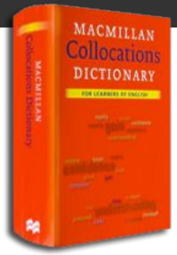


Mikolov et al. (2013)

CWN: How it works

Train data manual compilation and disambiguation

Manual selection of a few dozens collocations

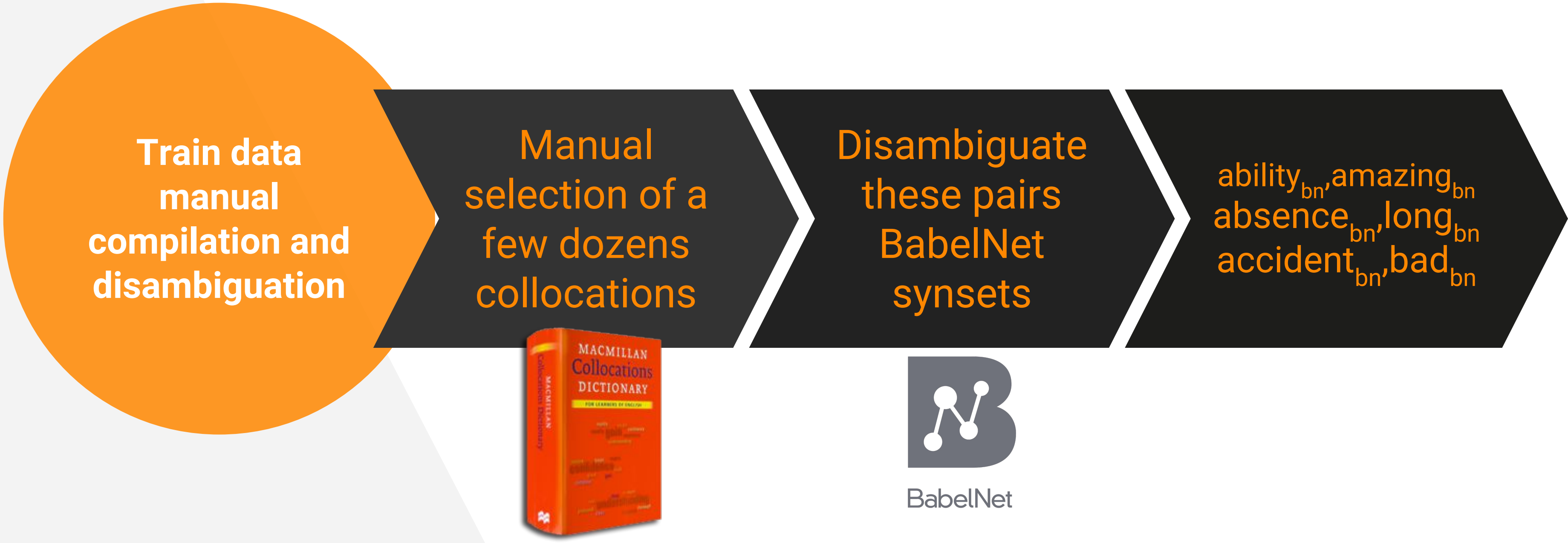


Disambiguate these pairs BabelNet synsets

BabelNet

ability,amazing
absence,long
accident,bad

CWN: How it works



Train data manual compilation and disambiguation

Manual selection of a few dozens collocations

Disambiguate these pairs BabelNet synsets

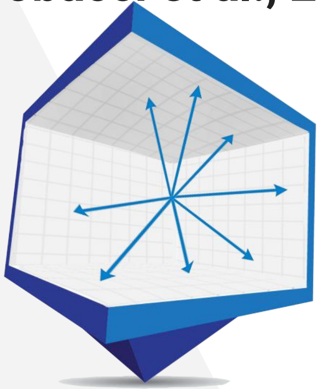
ability_{bn},amazing_{bn}
absence_{bn},long_{bn}
accident_{bn},bad_{bn}

CWN: How it works

Mapping train data sense-level embeddings models (vectors)

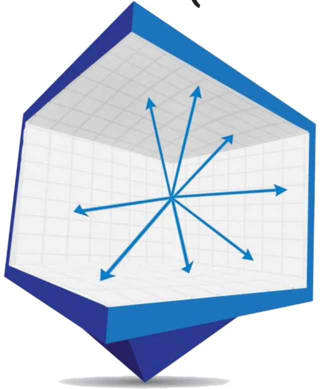
CWN: How it works

Bases (Iacobacci et al., 2015)

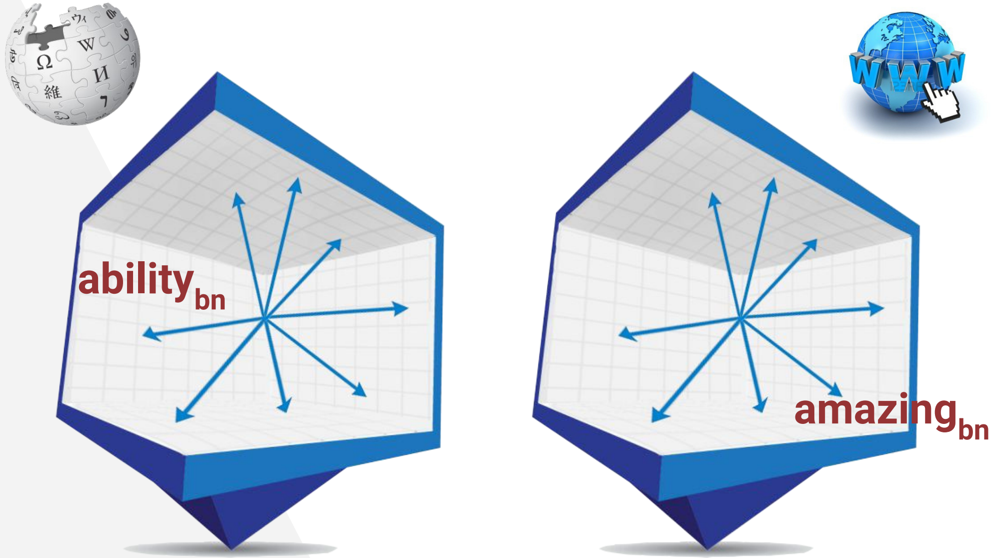


Mapping train data sense-level embeddings models

Collocates (Mancini et al., 2016)



CWN: How it works

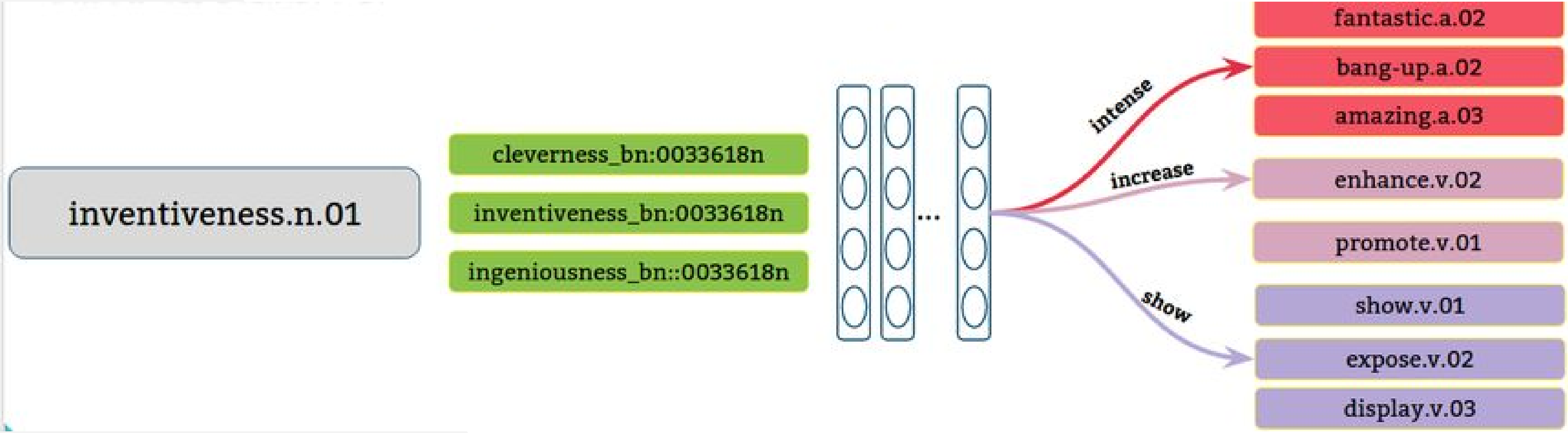


Train a transformation matrix between bases and collocates

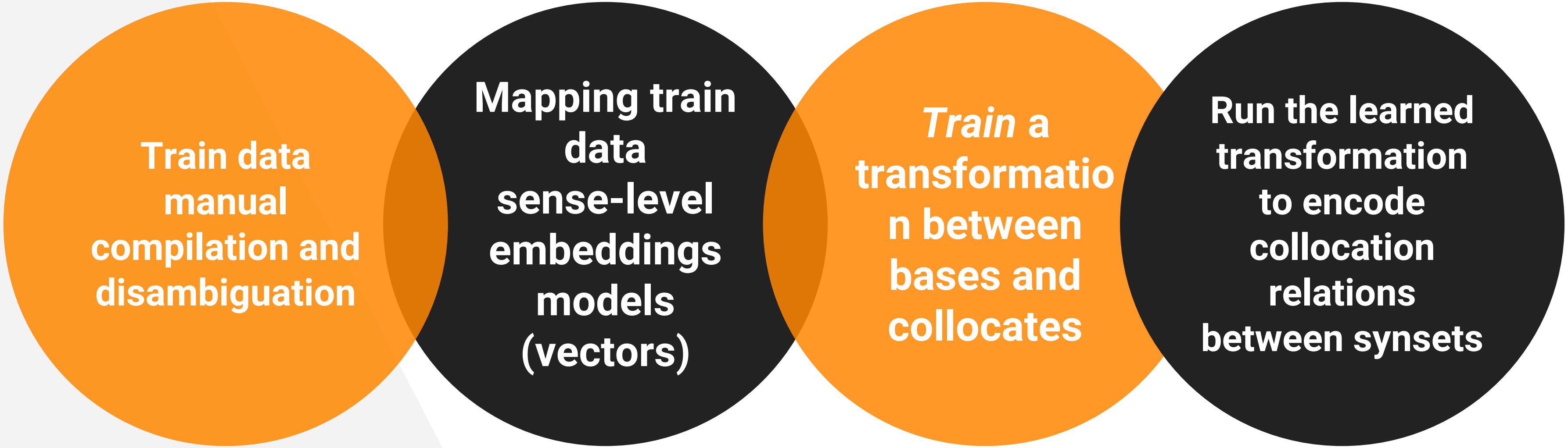
CWN: How it works

Run the learned transformation to encode collocation relations between synsets

CWN: How it works



CWN: How it works

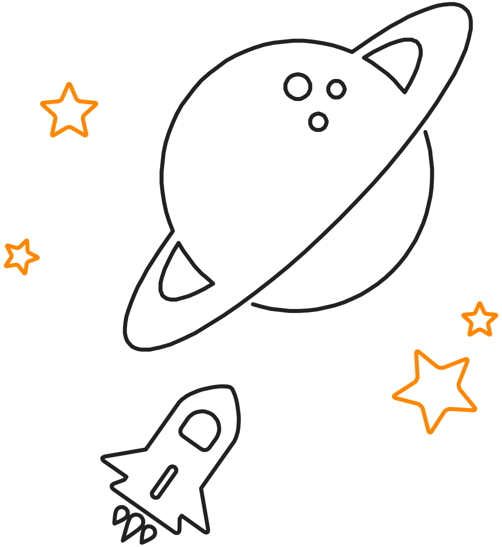


Train data manual compilation and disambiguation

Mapping train data sense-level embeddings models (vectors)

Train a transformation between bases and collocates

Run the learned transformation to encode collocation relations between synsets



The API

A preliminary approach to enabling CWN querying

What it is/does and what it isn’t/doesn’t

It generates a reliable mapping

The “translation matrix” approach has been validated in plenty of NLP tasks.

What it is/does and what it isn’t/doesn’t

It generates a reliable mapping

The “translation matrix” approach has been validated in plenty of NLP tasks.

It works equally well for all LFs

Dependent on the *amount* and *quality* of training data. New issues related to semantics that we cannot answer yet.

What it is/does and what it isn’t/doesn’t

It generates a reliable mapping

The “translation matrix” approach has been validated in plenty of NLP tasks.

It works equally well for all LFs

Dependent on the *amount* and *quality* of training data. New issues related to semantics that we cannot answer yet.

Multilingual potential

Since we used BabelNet as pivot, all assets provided by BabelNet can be explicitly leveraged.

What it is/does and what it isn’t/doesn’t

It generates a reliable mapping

The “translation matrix” approach has been validated in plenty of NLP tasks.

It is fast

It’s actually not very fast. We’ve experimented with additional transformation appraches, but the difference in speed and performance has not been evaluated yet.

It works equally well for all LFs

Dependent on the *amount* and *quality* of training data. New issues related to semantics that we cannot answer yet.

Multilingual potential

Since we used BabelNet as pivot, all assets provided by BabelNet can be explicitly leveraged.

What it is/does and what it isn’t/doesn’t

It generates a reliable mapping

The “translation matrix” approach has been validated in plenty of NLP tasks.

It works equally well for all LFs

Dependent on the *amount* and *quality* of training data. New issues related to semantics that we cannot answer yet.

Multilingual potential

Since we used BabelNet as pivot, all assets provided by BabelNet can be explicitly leveraged.

It is fast

It’s actually not very fast. We’ve experimented with additional transformation apprachеs, but the difference in speed and performance has not been evaluated yet.

It can be fine-tuned

Its algorithmic nature makes it possible to set specific thresholds for each lexical function.

What it is/does and what it isn’t/doesn’t

It generates a reliable mapping

The “translation matrix” approach has been validated in plenty of NLP tasks.

It works equally well for all LFs

Dependent on the *amount* and *quality* of training data. New issues related to semantics that we cannot answer yet.

Multilingual potential

Since we used BabelNet as pivot, all assets provided by BabelNet can be explicitly leveraged.

It is fast

It’s actually not very fast. We’ve experimented with additional transformation apprаches, but the difference in speed and performance has not been evaluated yet.

It can be fine-tuned

Its algorithmic nature makes it possible to set specific thresholds for each lexical function.

It is finished

This is a very preliminary prototype, so by no means this is a finished project.

Thank you

Any questions?

https://bitbucket.org/luisespinosa/cwn/

Twitter: @luisanke