

Multilinguality at Your Fingertips: BabelNet, Babelfy and Beyond!

Roberto Navigli

DIPARTIMENTO
DI INFORMATICA



SAPIENZA
UNIVERSITÀ DI ROMA

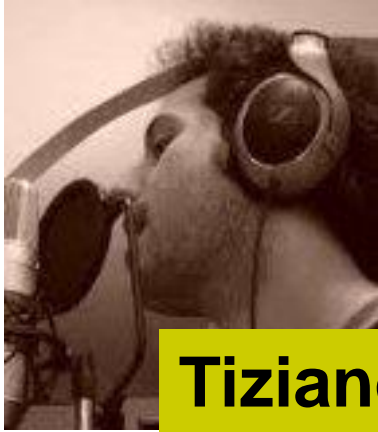
Linguistic Computing Laboratory

<http://lcl.uniroma1.it>

ERC Starting Grant n. 259234

LIDER CSA n. 610782

Moscow, 28th May 2015



**Tiziano
Flati**



**Daniele
Vannella**



**Andrea
Moro**



**Francesco
Cecconi**



**Taher
Pilehvar**

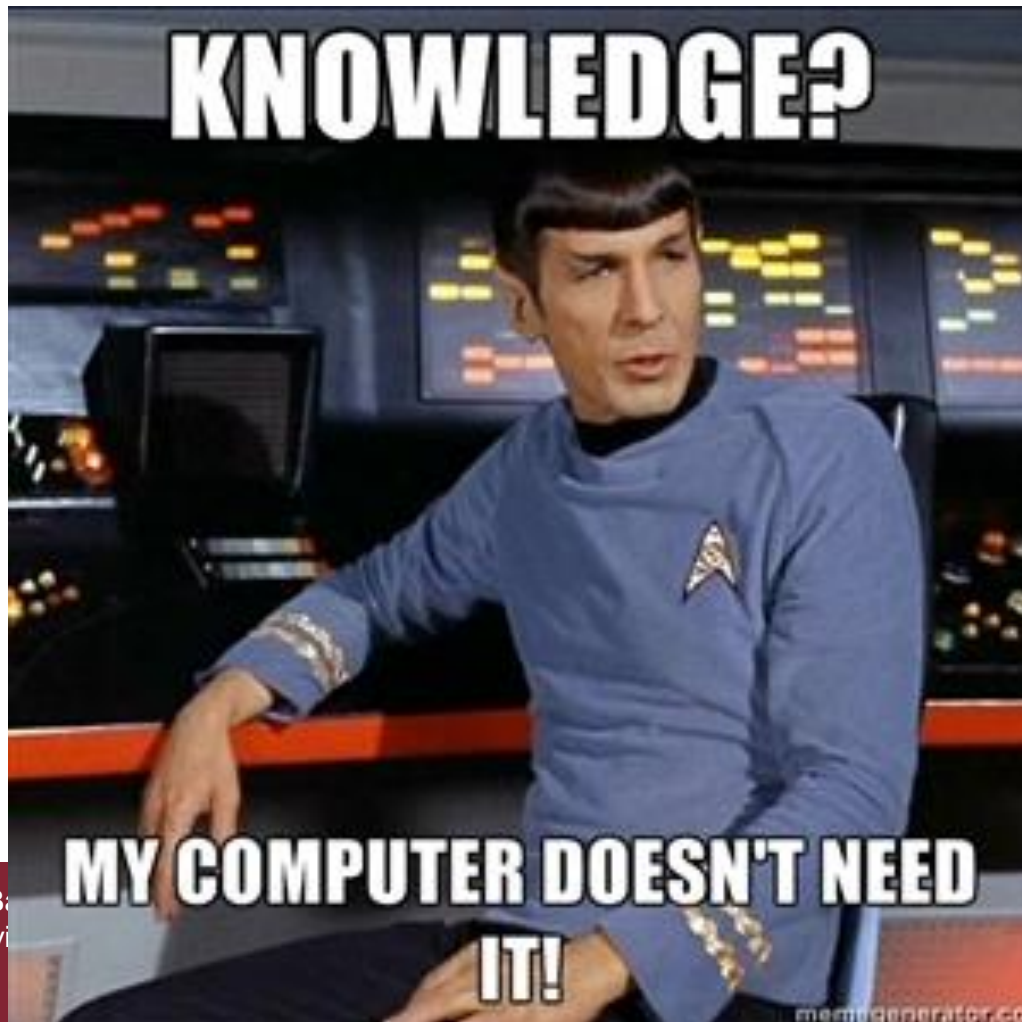


**Simone
Ponzetto**

knowledge

It's all about knowledge!

- But can we expect computers to **know**?
- Can't computers just use, e.g., **statistical techniques**?



State-of-the-art Machine Translation

- **EN:** These are movies in which the music genre, e.g. **rock**, is an important element but not necessarily central to the plot. Examples are Easy Rider (1969), The Graduate (1969), and Saturday Night Fever (1978).



State-of-the-art Machine Translation

- **EN:** These are movies in which the music genre, e.g. **rock**, is an important element but not necessarily central to the plot. Examples are Easy Rider (1969), The Graduate (1969), and Saturday Night Fever (1978).
- **FR:** Ce sont des films dans lesquels le genre de musique, par exemple, **rock**, est un élément important, mais pas nécessairement au centre de l'intrigue. Les exemples sont Easy Rider (1969), The Graduate (1969), et Saturday Night Fever (1978).

State-of-the-art Machine Translation

- **EN:** These are movies in which the music genre, e.g. **rock**, is an important element but not necessarily central to the plot. Examples are Easy Rider (1969), The Graduate (1969), and Saturday Night Fever (1978).
- **ES:** Estas son las películas en las que el género de la música, por ejemplo, **roca**, es un elemento importante, pero no necesariamente el centro de la trama. [...]



State-of-the-art Machine Translation

- **EN:** We can look at how this vast slug of molten underground **rock** was injected.

Danger here!



State-of-the-art Machine Translation

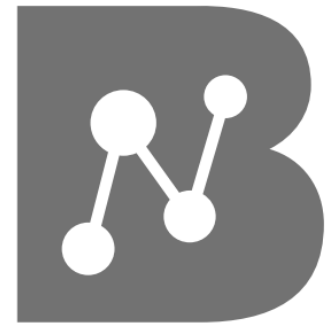
- **EN:** We can look at how this vast slug of molten underground **rock** was injected.
- **FR:** Nous pouvons voir comment ce vaste bouchon de **rock** underground fondu a été injecté.
- **IT:** Possiamo guardare a come è stato iniettato questo vasto slug del **rock** underground fusa.



What are we talking about?



A **5-year ERC Starting Grant** (2011-2016)
on Multilingual Word Sense Disambiguation



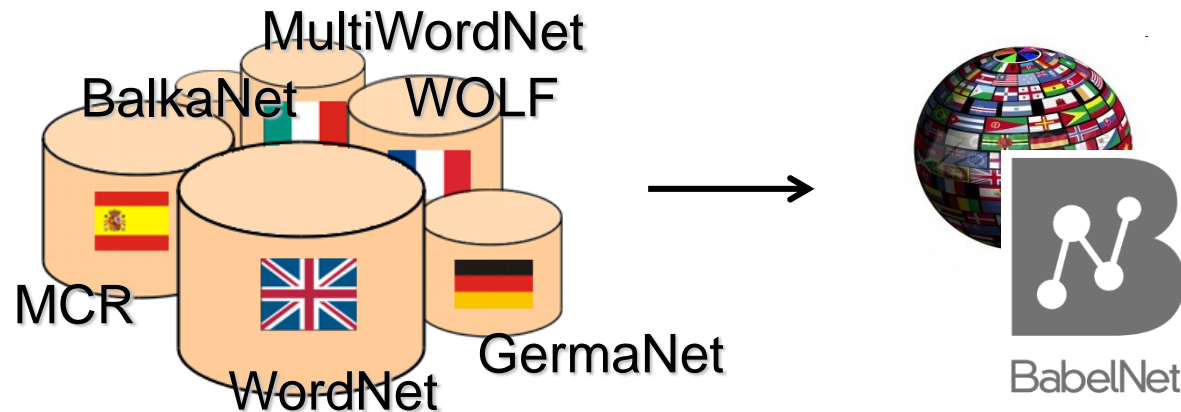
BabelNet

INTEGRATING KNOWLEDGE

[Navigli & Ponzetto, ACL 2010;
Pilehvar & Navigli, ACL 2014]

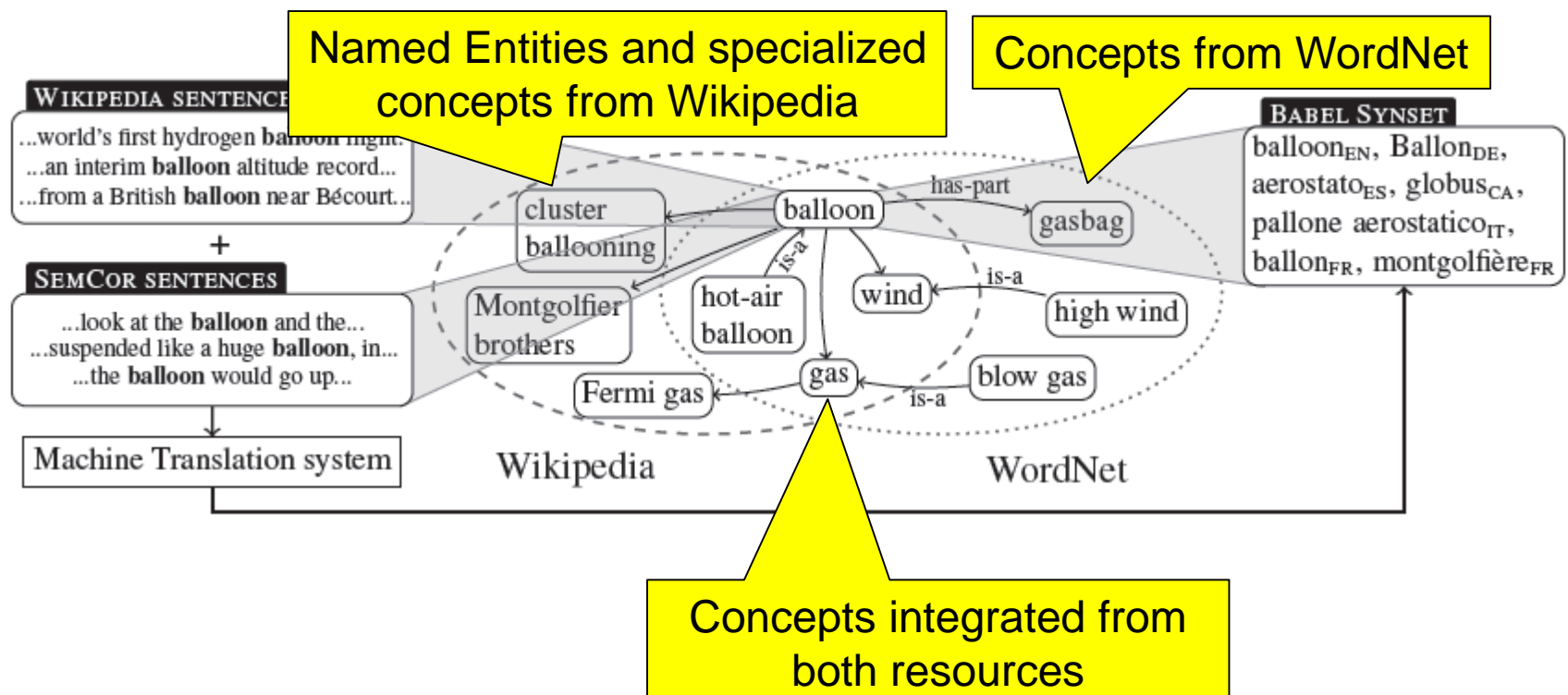
Multilingual Joint Word Sense Disambiguation (MultiJEDI)

Key Objective 1: create **knowledge** for **all languages**



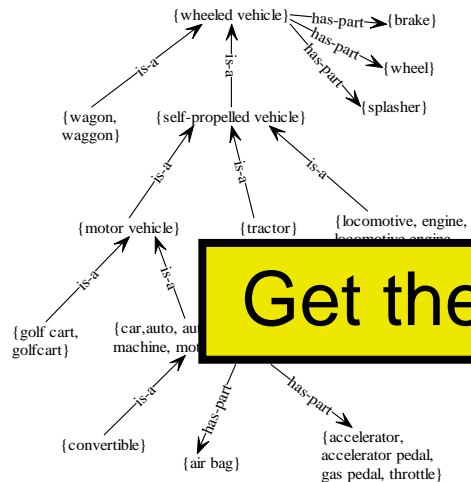
It all started with merging WordNet and Wikipedia [Navigli and Ponzetto, ACL 2010; AIJ 2012]

- A wide-coverage multilingual semantic network including both **encyclopedic** (from Wikipedia) and **lexicographic** (from WordNet) entries

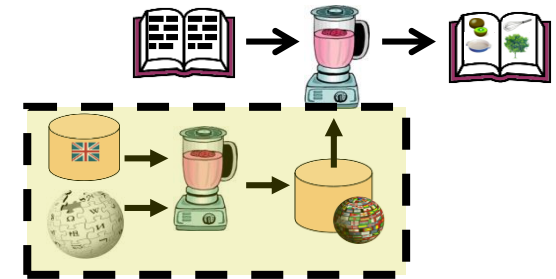


Creating a Multilingual Semantic Network

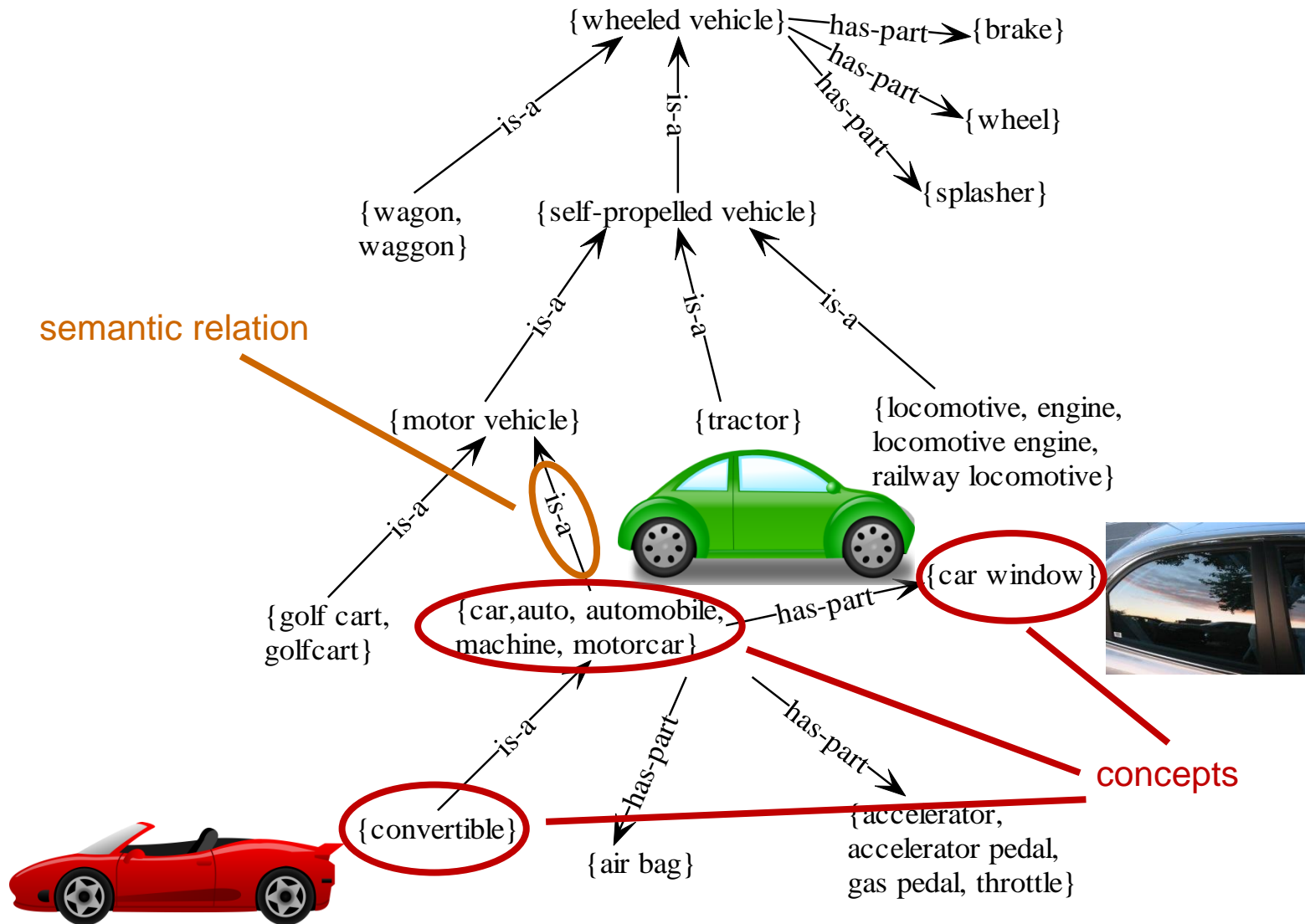
- Start from two large **complementary** resources:
 - WordNet**: full-fledged taxonomy
 - Wikipedia**: multilingual and continuously updated



Get the best from both worlds



WordNet [Miller et al., 1990; Fellbaum, 1998]



Wikipedia [The Web Community, 2001-today]

Automobile

From Wikipedia, the free encyclopedia
(Redirected from [Car](#))

For the magazine, see [Automobile Magazine](#).

"[Car](#)" redirects here. For other uses, see [Car \(disambiguation\)](#).

An **automobile**, **autocar**, **motor car** or **car** is a wheeled [motor vehicle](#) used for [transporting passengers](#), which also carries its own engine or motor. Most definitions of the term specify that automobiles are designed to run primarily on roads, to have seating for one to eight people, to typically have four wheels, and to be constructed principally for the transport of people rather than goods.^[3]

The term *motorcar* has also been used in the context of electrified rail systems to denote a car which functions as a small locomotive but also provides space for passengers and baggage. These locomotive cars were often used on suburban routes by both interurban and intercity railroad systems.^[4]

It was estimated in 2010 that the number of automobiles had risen to over 1 billion vehicles, with 500 million reached in 1986.^[5] The numbers are increasing rapidly, especially in [China](#) and [India](#).^[6]



(unspecified) semantic relation

concepts

Passenger

From Wikipedia, the free encyclopedia

This article is about passengers in commercial transportation; for other uses see [Passenger \(disambiguation\)](#)

A **passenger** is a person who travels in a [vehicle](#) but bears little or no responsibility for the tasks required for that vehicle to arrive at its destination or otherwise operate the vehicle.

Passengers are people who ride on [buses](#), [passenger trains](#), [airliners](#), [ships](#), [ferryboats](#), and other methods of transportation.

Crew members (if any), as well as the driver or pilot of the vehicle, are considered to be passengers. For example, a [flight attendant](#) would not be considered a "passenger" while on duty, but an [idling](#) in a [company car](#) being driven by another person would be a *passenger*, even if the car was being driven in company

Look up [passenger](#) in Wiktionary, the free dictionary.



Motor vehicle

From Wikipedia, the free encyclopedia

A **motor vehicle** or **road vehicle** is a self-propelled wheeled [vehicle](#) that does not operate on rails, such as [trains](#) or [trolleys](#). The [vehicle propulsion](#) is provided by an [engine](#) or motor, usually by an [internal combustion engine](#), or an [electric motor](#), or some combination of the two, such as [hybrid electric vehicles](#) and [plug-in hybrids](#). For legal purposes motor vehicles are often identified within a number of vehicle classes including [automobiles](#) or cars, [buses](#), [motorcycles](#), [motorized bicycles](#), [off highway vehicles](#), [light trucks](#) or light duty trucks, and [trucks](#) or lorries. These classifications vary according to the legal codes of each country. ISO 3833:1977 is the standard for road vehicles types, terms and definitions.^[1]

As of 2010 there were more than one billion motor vehicles in use in the world excluding [off-road vehicles](#) and [heavy construction equipment](#).^{[2][3][4]} Global vehicle ownership [per capita](#) in 2010 was 148 vehicles in operation per 1000 people.^[4] The United States has the largest fleet of motor vehicles in the world, with 239.8 million by 2010. Vehicle



The [United States](#) has the world's largest motor vehicle registered fleet, with almost 250 million vehicles.

[hide](#)

Travel

From Wikipedia, the free encyclopedia
(Redirected from [Traveling](#))

For other uses, see [Travel \(disambiguation\)](#).

Travel is the movement of [people](#) or objects (such as [airplanes](#), [boats](#), [trains](#) and other conveyances) between relatively distant geographical [locations](#).^{[1][2]}

Contents [\[hide\]](#)

- [1 Etymology](#)
- [2 Purpose and motivation](#)
- [3 Travel safety](#)
- [4 See also](#)
- [5 References](#)
- [6 External links](#)



A statue dedicated to the traveler in Oviedo, Spain.

Etymology

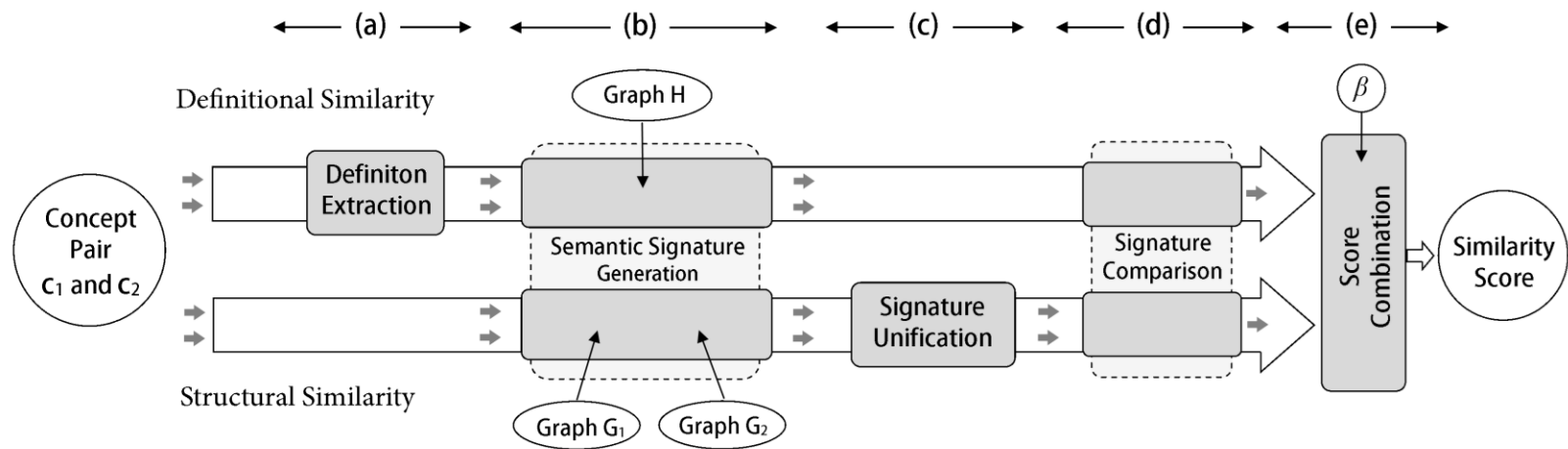
The term "travel" originates from the Old French word *travail*.^[3] The term also covers all the activities performed during a travel (movement).^[4]

Roberto Navigli

SemAlign: Cross-resource Concept Alignment

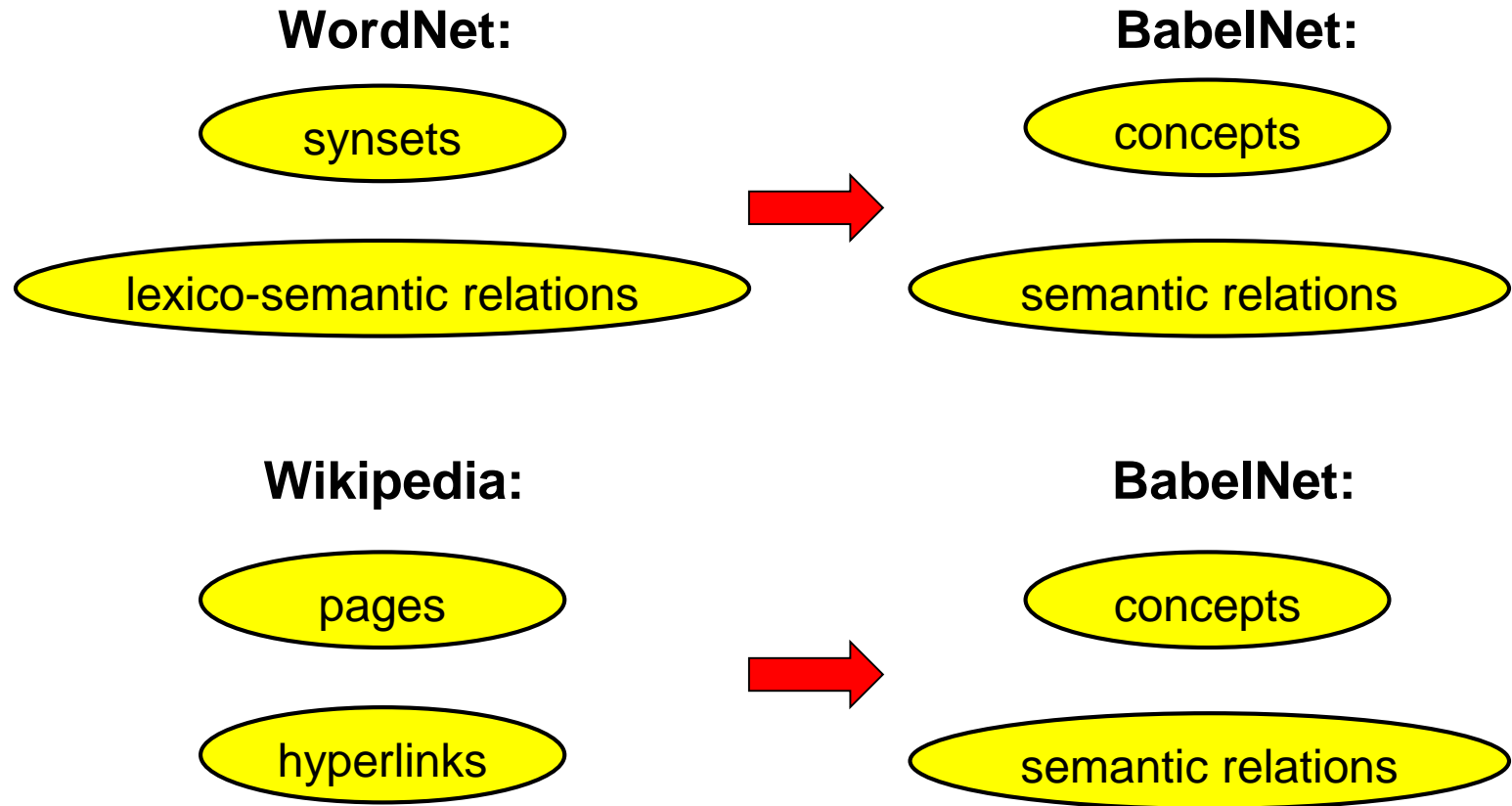
[Pilehvar and Navigli, ACL 2014]

We combine two different similarity measures:



BabelNet: concepts and semantic relations (1)

- **Concepts** and **relations** in BabelNet are harvested from the various resources:



BabelNet: concepts and semantic relations (2)

- We encode knowledge as a **labeled directed graph**:
 - Each vertex is a **Babel synset**



- Each edge is a **semantic relation** between synsets:
 - **is-a** (balloon is-a aircraft)
 - **part-of** (gasbag part-of balloon)
 - **instance-of** (Einstein instance-of physicist)
 - ...
 - **unspecified/relatedness** (balloon related-to flight)

Building BabelNet: Translating Babel synsets

1. Exploiting Wikipedia interlanguage links

Ballon

globo
aerostático

pallone
aerostatico

About Wikipedia

Community portal

Recent changes

Contact Wikipedia

► Toolbox

► Print/export

▼ Languages

Ænglisc

العربية

Català

Česky

Cymraeg

Dansk

Deutsch

Eesti

Ελληνικά

Español

Esperanto

فارسی

Français

Frysk

한국어

Hrvatski

Bahasa Indonesia

Íslenska

Italiano

עברית

Қазақша

Lietuvių

മലയാളം

日本語

Norsk (bokmål)

Polski

Português

Română

Русский

automatic equipment (including cameras and [telescopes](#), and night-control mechanisms) may also be called the gondola.

Contents [hide]

1 Types

2 History

3 As flying machines

4 Military use

4.1 American Civil War

4.2 After the American Civil War

5 Records

6 In space

7 Sports

8 See also

9 References

10 External links


Types [edit]

There are three main types of balloons:

- [hot air balloons](#) obtain their buoyancy by heating the air inside the balloon. They are the most common type of balloon aircraft. "Hot air balloon" is sometimes used incorrectly to denote any balloon that carries people.
- [gas balloons](#) are inflated with a gas of lower [molecular weight](#) than the ambient atmosphere. Most gas balloons operate with the internal pressure of the gas the same as the [pressure of the surrounding atmosphere](#). There is a type of gas balloon, called a [superpressure balloon](#), that can operate with the [lifting gas](#) at pressure that exceeds the pressure of the surrounding air, with the objective of limiting or eliminating the loss of gas from day-time heating. Gas balloons are filled with gases such as:
 - [hydrogen](#) – not widely used for aircraft since the [Hindenburg disaster](#) because of high flammability (except for some sport balloons as well as nearly all unmanned scientific and weather balloons).
 - [helium](#) – the gas used today for all airships and most manned balloons.
 - [ammonia](#) – used infrequently due to its caustic qualities and limited lift.
 - [coal gas](#) – used in the early days of ballooning; it is highly flammable.
 - [methane](#) – used as a lower cost lifting gas, but offering less lift than helium or hydrogen.^[1]
- [Rozière balloons](#) use both heated and unheated lifting gases. The most common modern use of this type of balloon is for long-distance record flights such as the [recent circumnavigations](#).

History [edit]

Main article: [History of ballooning](#)



Building BabelNet: Translating Babel synsets

2. Filling the **lexical translation gaps** using a **Machine Translation** system to **translate** the English lexicalizations of a concept

- On August 27, 1783 in Paris, Franklin witnessed the world's first hydrogen **[[Balloon (aircraft)|balloon]]** flight.

Statistical Machine Translation

- Le 27 Août, 1783 à Paris, Franklin vu le premier vol en **ballon** d'hydrogène.

The most frequent translation of a word in a given meaning

| left context | term | right context |
|--------------------------------|--------------|--|
| | wikification | may refer to: the... |
| geoinformatics services' and ' | wikification | of GIS by the masses' |
| the process may be called | wikification | (as in ... |
| which is then called " | wikification | and to the related problem |
| reason needs copyediting, | wikification | , reduction of POV, work on references |
| huge amount of cleanup, | wikification | , etc. Version of 12 Nov |

The most frequent translation of a word in a given meaning

| left context | term | right context |
|-------------------------------|---------------|---|
| | wikificazione | potrebbe riferirsi a: il... |
| servizi geoinformatici' e ' | wikification | di GIS dalle masse' |
| il processo chiamato | wikificazione | (come in ... |
| che è quindi chiamato | wikificazione | e al problema correlato... |
| ragione richiede copyediting, | wikification | , riduzione di POV, lavoro su reference |
| grandi quantità di pulizia, | wikificazione | , ecc. Versione del 12 Novembre |

The most frequent translation of a word in a given meaning

| left context | term | right context |
|-------------------------------|---------------|---|
| | wikificazione | potrebbe riferirsi a: il... |
| servizi geoinformatici' e ' | wikification | di GIS dalle masse' |
| il processo chiamato | wikificazione | (come in ... |
| che è quindi chiamato | wikificazione | e al problema correlato... |
| ragione richiede copyediting, | wikification | , riduzione di POV, lavoro su reference |
| grandi quantità di pulizia, | wikificazione | , ecc. Versione del 12 Novembre |

What is BabelNet?

- A **merger** of resources of different kinds:

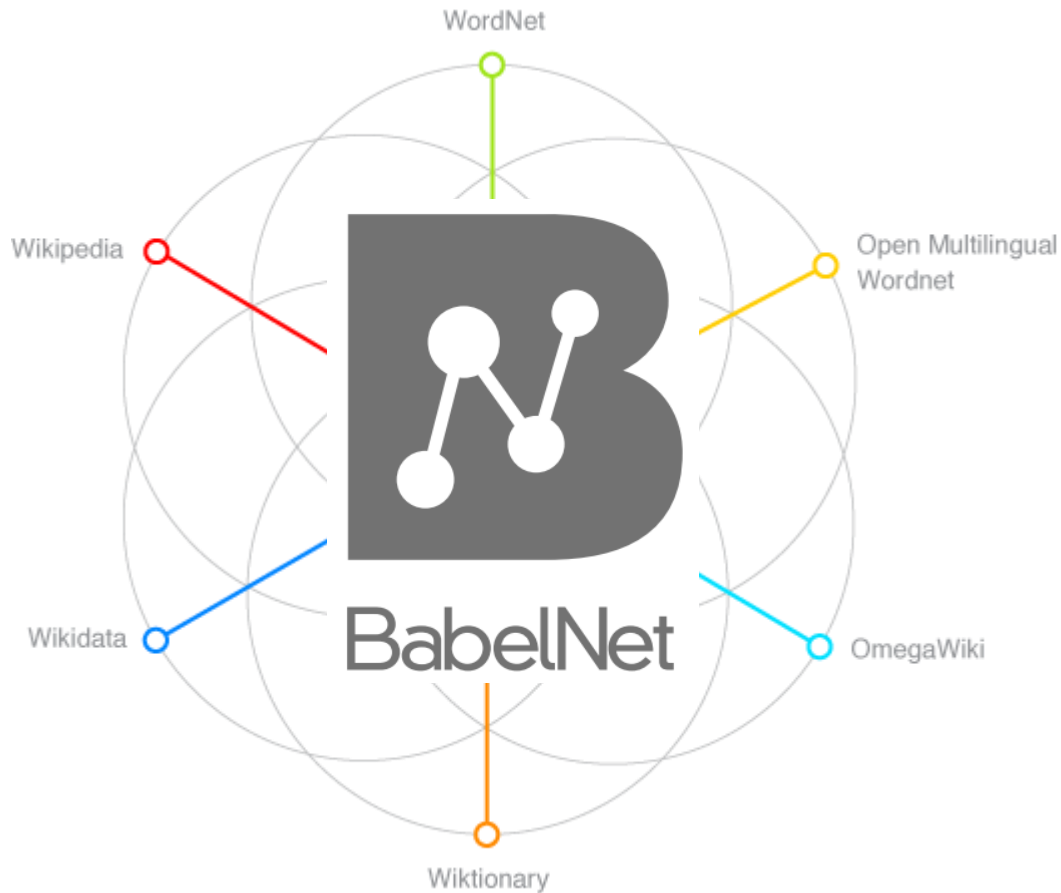


What is BabelNet?

- A **merger** of resources of different kinds:
 - **WordNet**: the most popular computational lexicon of English
 - **Open Multilingual WordNet**: a collection of open wordnets
 - **Wikipedia**: the largest collaborative encyclopedia
 - **Wikidata**: the largest collaborative knowledge base
 - **Wiktionary**: the largest collaborative dictionary
 - **OmegaWiki**: a medium-size collaborative multilingual dictionary
 - High-quality automatic **sense-based translations**

What is BabelNet?

- A **merger** of resources of different kinds:




Why do we need BabelNet?

- **Multilinguality**: the same concept is expressed in tens of languages

- Dictionary
- Images
- Translations
- Sources
- Categories
- External links

EnglishArabicChineseFrenchGermanGreekHebrewHindiItalianJapanese+ all preferred languages







bn:00002838n • NOUN • Concept • Categories: Bicycle tools, Mechanical hand tools, Screws


 **Allen wrench** • Hex key

A wrench for Allen screws + More definitions

IS-A: wrench • tool • hand tool

EXPLORE NETWORK





Why do we need BabelNet?

- **Multilinguality**: the same concept is expressed in tens of languages

Translations

-  مفك سداسي, مفتاح سداسي, ألن وجع, وجع ألين, مفتاح ألين, مفتاح عرافة
-  内六角扳手, 六角匙, 内六角扳手, 内六角扳手, 六角 扳 手
-  Allen wrench, Hex key, Allen key, Hex head wrench, Allen bolt, Allan keys, Inbus, Alum key, Allan wrench, Zeta key, Allen socket, Hex wrench, Allum key, Unbrako, Alan wrench, Alan key, Allen keys, Imbus, Hex driver, Allan key, Socket head, Umbrako
-  Clé Allen, clef Allen, Clef six pans, Clé six pans creux, *clé hexagonale*
-  Inbusschlüssel, Innensechskantschlüssel, Innensechskant, Inbus, Inbusschraube, Innensechskantschraube Bauer und Schaurte, Sechskantschlüssel, Innensechskantschraube, Sechskantschraubendreher
-  κλειδί allen, εξαγωνα κλειδί
-  מפתח אלן, אלן מפתח ברגים, מפתח ברגים, אלן מפתח
-  एलन रिंग, हेक्स कुंजी
-  Chiave a brugola, Brugola, Viti brugola, Imbus, Chiave di Allen, Chiave Allen, *chiave esagonale*
-  六角棒スパナ, 六角レンチ, 六角棒レンチ, ヘキサゴンレンチ, アーレンキー, 六角レンチ。
-  Шестигранный ключ, Шестигранный шлиц, Инбусовый ключ, Инбус, Имбусовый ключ, Шестигранник
-  llave allen, Llaves allen, Tornillo allen, *llave hexagonal*

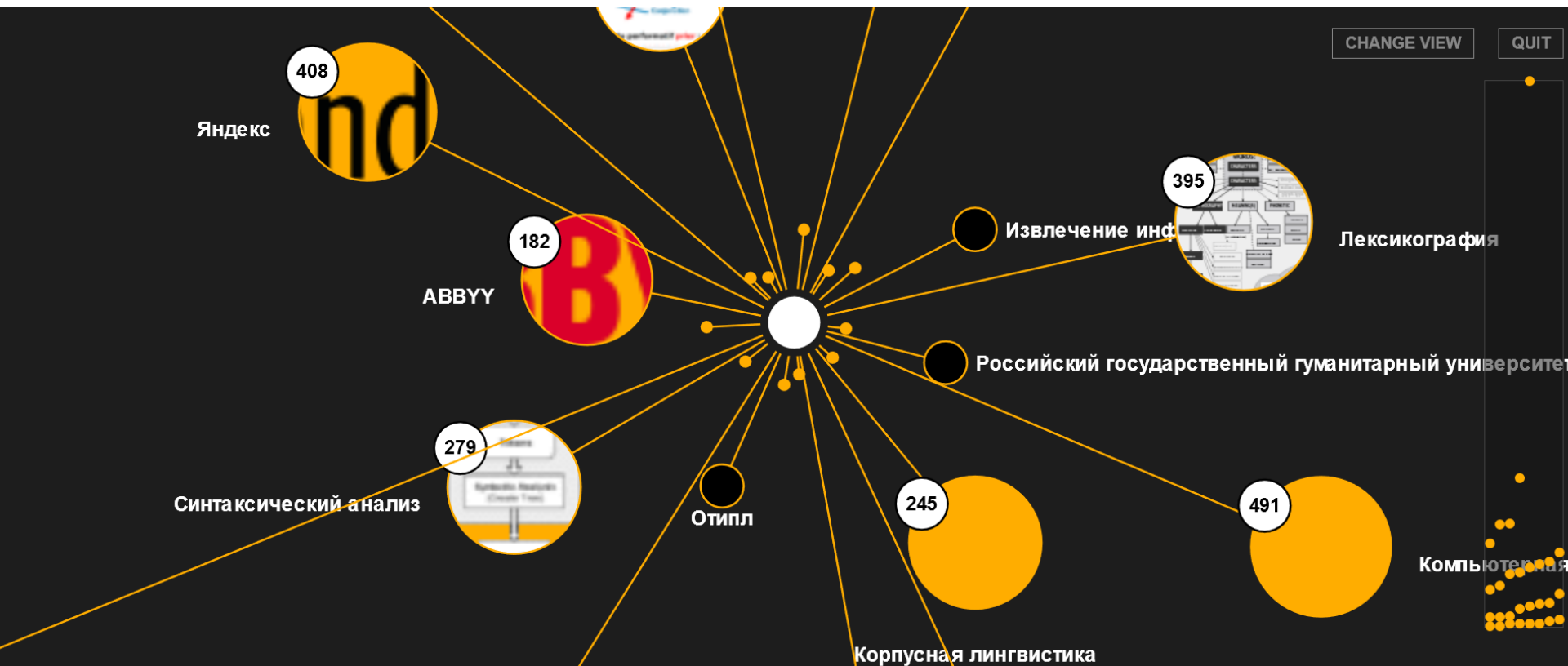
Why do we need BabelNet?

- **Multilinguality**: the same concept is expressed in tens of languages
- **Coverage**: 271 languages and 14 million entries!



Why do we need BabelNet?

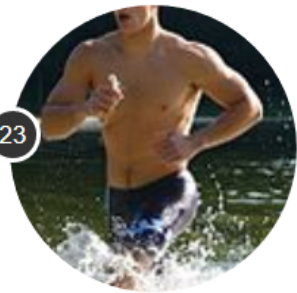
- **Multilinguality**: the same concept is expressed in tens of languages
- **Coverage**: 271 languages and 14 million entries!
- **Concepts and named entities together**: dictionary and encyclopedic knowledge is semantically interconnected



Why do we need BabelNet?

- **Multilinguality**: the same concept is expressed in tens of languages
- **Coverage**: 271 languages and 14 million entries!
- **Concepts and named entities together**: dictionary and encyclopedic knowledge is semantically interconnected
- **"Dictionary of the future"**: semantic network structure with labeled relations, pictures, multilingual synsets

Verb



run

Move fast by using one's feet, with one foot off the ground at any given time

ID: 00093170v | Concept

هَرُؤَلْ, جَرَى, رَغَضَن

奔跑, 跑

courir

rennen

τρέχω, κινούμαι

ץר

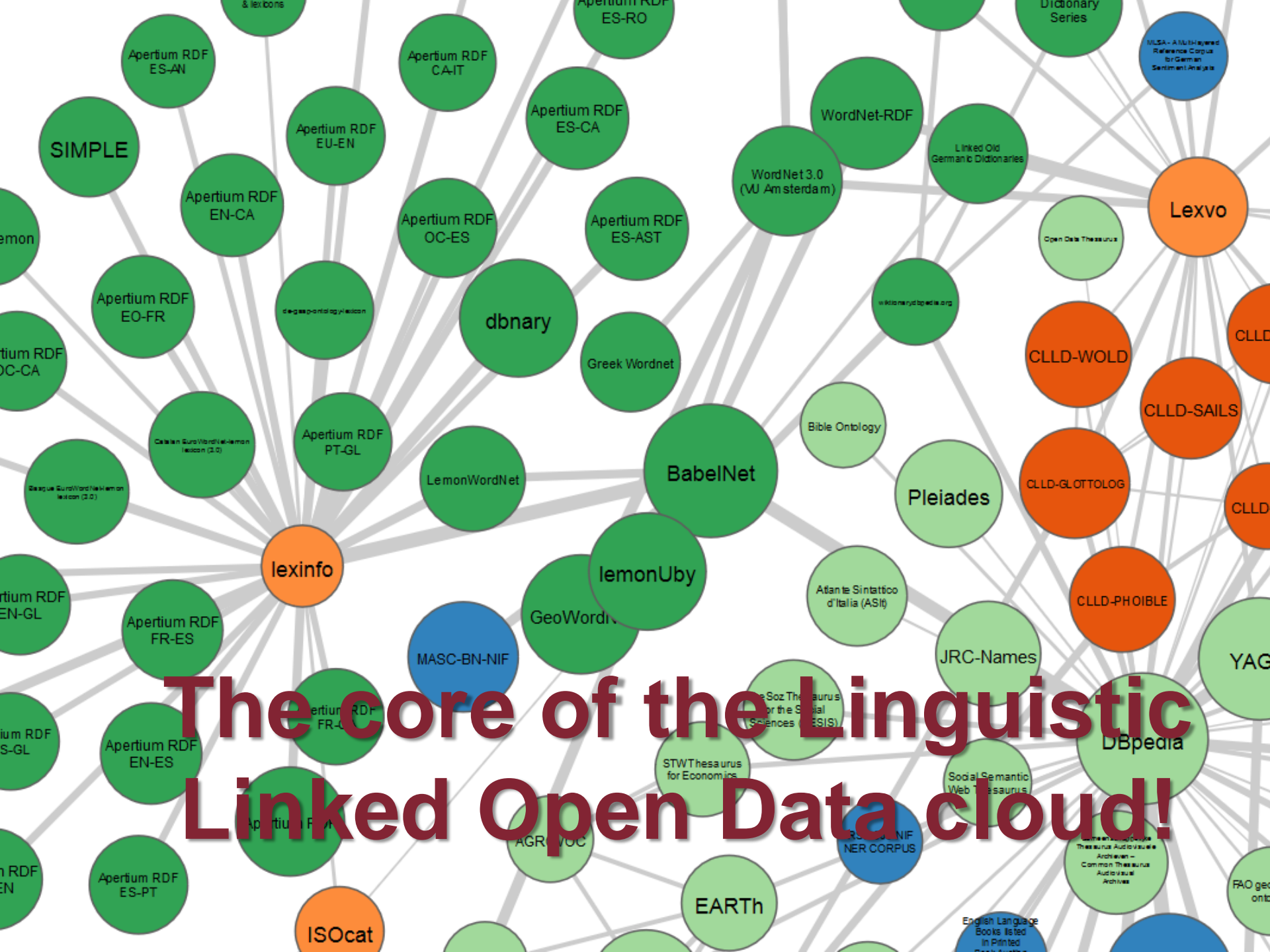
correre

Why do we need BabelNet?

- **Multilinguality**: the same concept is expressed in tens of languages
- **Coverage**: 271 languages and 14 million entries!
- **Concepts and named entities together**: dictionary and encyclopedic knowledge is semantically interconnected
- **"Dictionary of the future"**: semantic network structure with labeled relations, pictures, multilingual synsets
- **Full-fledged taxonomy**: is-a relations are available for both concepts and named entities (Wikipedia Bitaxonomy)
 - ABBYY *is-a* software company, company
 - Tim Berners-Lee *is-a* computer scientist

Why do we need BabelNet?

- **Multilinguality**: the same concept is expressed in tens of languages
- **Coverage**: 271 languages and 14 million entries!
- **Concepts and named entities together**: dictionary and encyclopedic knowledge is semantically interconnected
- **"Dictionary of the future"**: semantic network structure with labeled relations, pictures, multilingual synsets
- **Full-fledged taxonomy**: is-a relations are available for both concepts and named entities (Wikipedia Bitaxonomy)
 - ABBYY *is-a* software company, company
 - Tim Berners-Lee *is-a* computer scientist
- **Easy access**: Java and HTTP RESTful APIs; SPARQL endpoint (2 billion triples)



What can we do with BabelNet?

- Search and translate:

The screenshot displays the BabelNet website. At the top center is the BabelNet logo, a stylized 'B' with a network diagram inside, and the text 'BabelNet' below it. A tagline reads: 'A very large multilingual encyclopedic dictionary and semantic network'. In the top right corner, there are links for 'LOG IN' and 'REGISTER'. The main interface features a search bar with the word 'plane' entered. To the right of the search bar are dropdown menus for 'ENGLISH' and '3 SELECTED'. A teal 'TRANSLATE' button is on the far right. Below the search bar, a teal banner states: 'THE BABELNET 3.0 JAVA & HTTP APIS ARE AVAILABLE NOW'. A 'PREFERENCES' link with a gear icon is visible. A language selection dropdown is open, showing a list of languages with checkboxes: ARABIC, CHINESE (checked), FRENCH, GERMAN (checked), GREEK, HEBREW, HINDI, and ITALIAN (checked). The footer contains a small BabelNet logo, a list of links (ABOUT, PUBLICATIONS, STATS, DOWNLOADS, API GUIDE), a paragraph of text about the project's funding and license, and logos for 'STUDIVM PARIS' and 'erc'.

plane

ENGLISH

3 SELECTED

TRANSLATE

Search

THE BABELNET 3.0 JAVA & HTTP APIS ARE AVAILABLE NOW

⚙️ PREFERENCES

☐ ARABIC

☒ CHINESE

☐ FRENCH

☒ GERMAN

☐ GREEK

☐ HEBREW

☐ HINDI

☒ ITALIAN

ABOUT
PUBLICATIONS
STATS
DOWNLOADS
API GUIDE

BabelNet is an output of the [MultiJEDI ERC Starting Grant](#) No. 259234. Concept and application by [Roberto Navigli](#). BabelNet and its API are licensed under a [Creative Commons Attribution-Non Commercial-Share Alike 3.0 License](#). For any commercial use, please [contact us](#).

STUDIVM PARIS

erc

What can we do with BabelNet?

- Noun
- Verb
- Adjective

Noun



airplane, plane, aeroplane

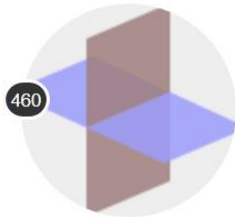
An aircraft that has a fixed wing and is powered by propellers or jets

ID: 00001697n | Concept

固定翼飛機, 飛行機, 飞龙机

Flugzeug

aereo, aeroplano, apparecchio



plane, sheet

(mathematics) an unbounded two-dimensional shape

ID: 00062766n | Concept

平面, 面

Ebene (Mathematik)

piano, piano geometrico



plane

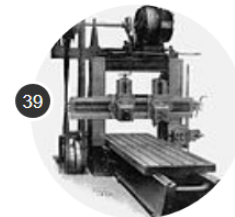
A level of existence or development

ID: 00062767n | Concept

平面的存在

Ebene

piano, Spostamento della realtà, livello



planer, plane, planing machine

A power tool for smoothing or shaping wood

ID: 00062768n | Concept

刨床

Hobelmaschine

piallatrice



plane, woodworking plane, carpenter's plane

A carpenter's hand tool with an adjustable blade for smoothing or shaping wood

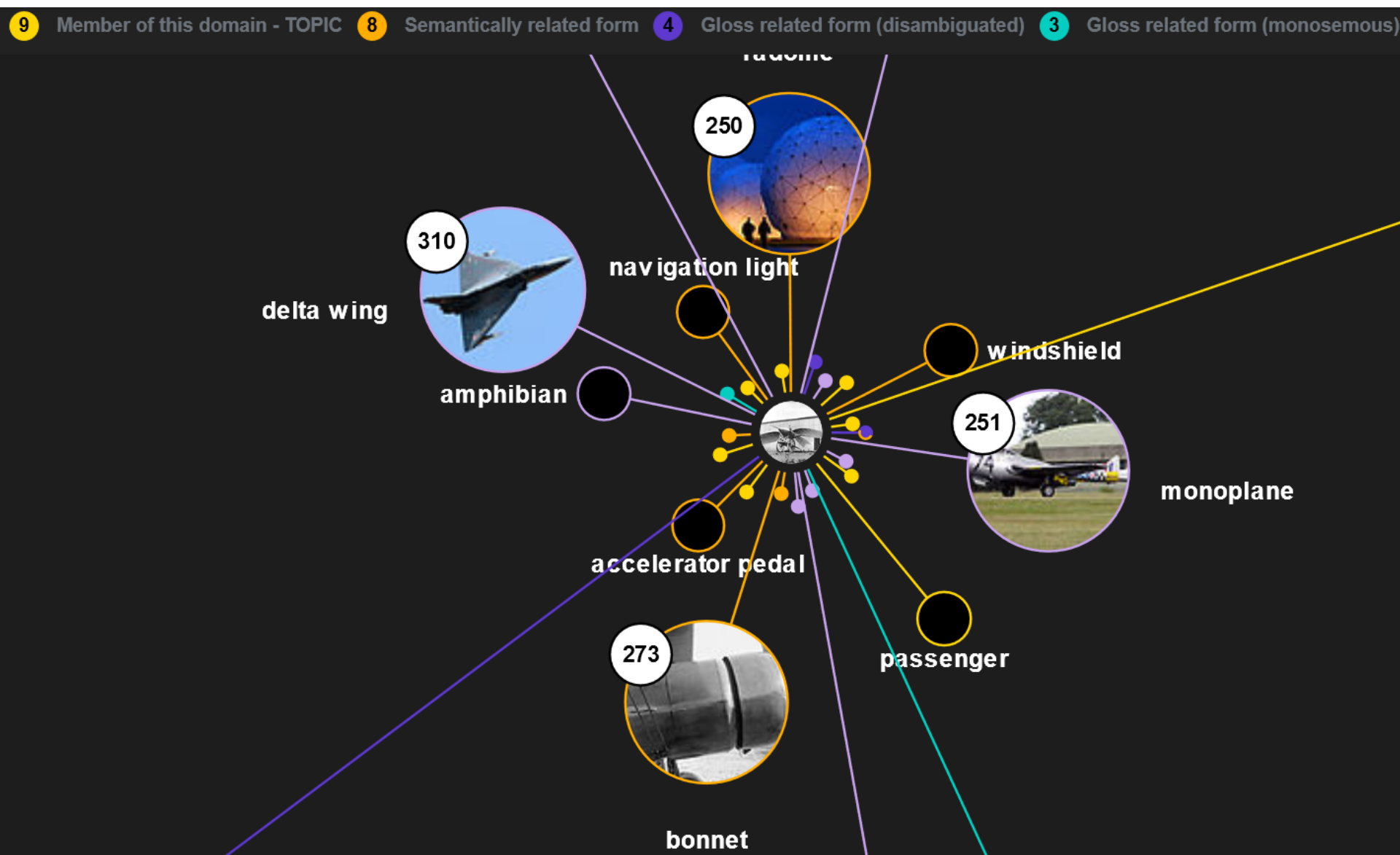
刨

Hobel

pialla, piana, pialletto

What can we do with BabelNet?


- Explore the network:



WordNet-Wikipedia mapping accuracy

- Quality **lower bound** of the mapping: **87%**
 - On the 6000 lowest-confidence mappings
 - Note: this **concerns only** 50k synsets in the intersection



A full-page photograph of a night sky filled with stars and the Milky Way galaxy. In the lower right foreground, the dark silhouette of a person stands with their back to the camera, looking up at the vast cosmic display. The horizon is a dark, slightly uneven line at the bottom of the frame.

*“Interestingly, the feeling of being all alone
in the entire Universe can be
mystically beautiful”*

We are not alone in the (resource) universe!

- **DBPedia** [Bizer et al. 2009] - a resource obtained from structured information in Wikipedia
 - «Describes 3.77M things»
 - No dictionary side
- **YAGO** [Suchanek et al. 2007]
 - «Contains 10M entities and 120M facts about these entities»
 - Links Wikipedia categories to WordNet synsets
- **MENTA** [de Melo and Weikum, 2010]
 - A «multilingual taxonomy with 5.4M entities»
- **WikiNet** [Nastase and Strube, 2013]
 - Semantic network connecting Wikipedia entities
 - «3M concepts and 38+M relations»
- **Freebase** (<http://freebase.com>): collaborative effort
 - Started from Wikipedia, MusicBrainz, ChefMoz, etc. **Shut down!**

Key fact!

Annotating with BabelNet:
all in one!

- Annotating with **BabelNet** implies annotating with WordNet, Wikipedia, OmegaWiki, Open Multilingual WordNet **and** Wiktionary



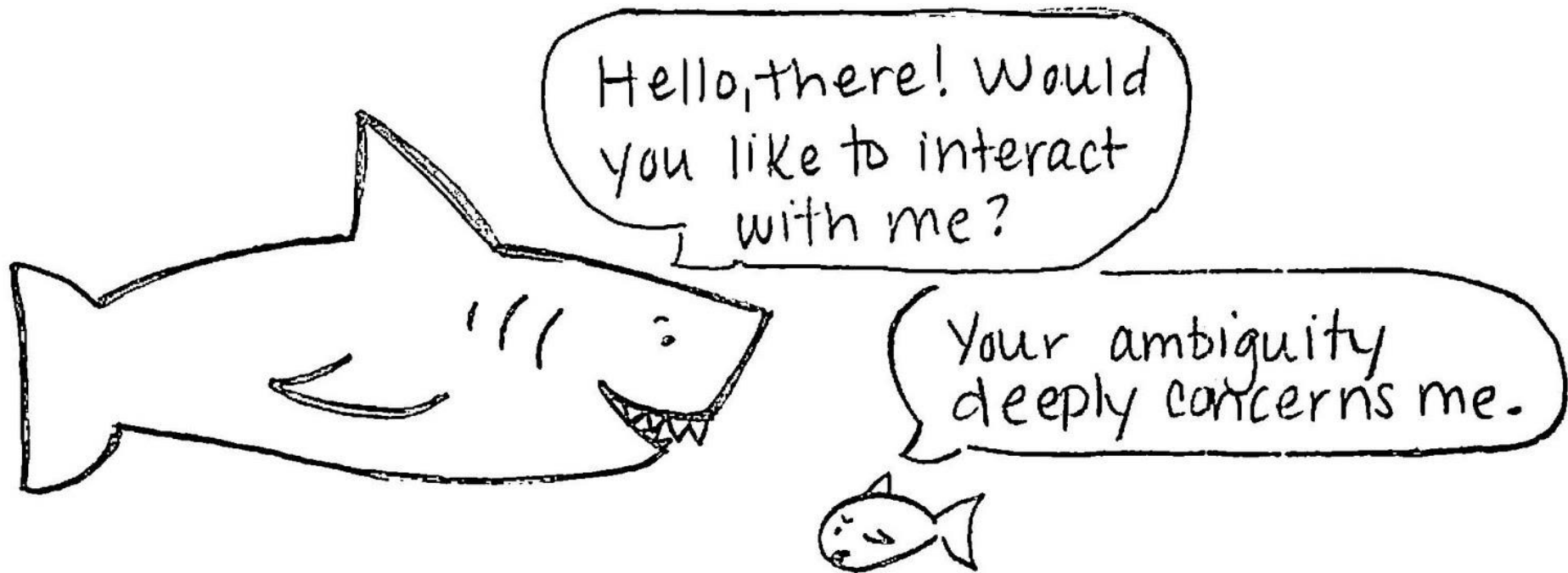
ONE DOES NOT SIMPLY USE BABELNET

AS A MULTILINGUAL DICTIONARY

ADDRESSING AMBIGUITY

[Moro, Raganato & Navigli,
TACL 2014]

Context matters!!!



Back to our issue: lexical ambiguity!

- Thomas and Mario played as strikers in Munich.

Thomas

and

Mario

played

as

strikers

in

Munich



Thomas

Thomas Müller is a German footballer who plays for Bayern Munich and the



Mario

Mario Gómez García is a German footballer who plays as a striker for Bayern Munich in

played

participate in games or sport; "We played hockey all afternoon"; "play cards"; "Pele



strikers

a forward on a soccer team



Munich

FC Bayern Munich, is a German sports club based in Munich, Bavaria.

Word Sense Disambiguation and Entity Linking

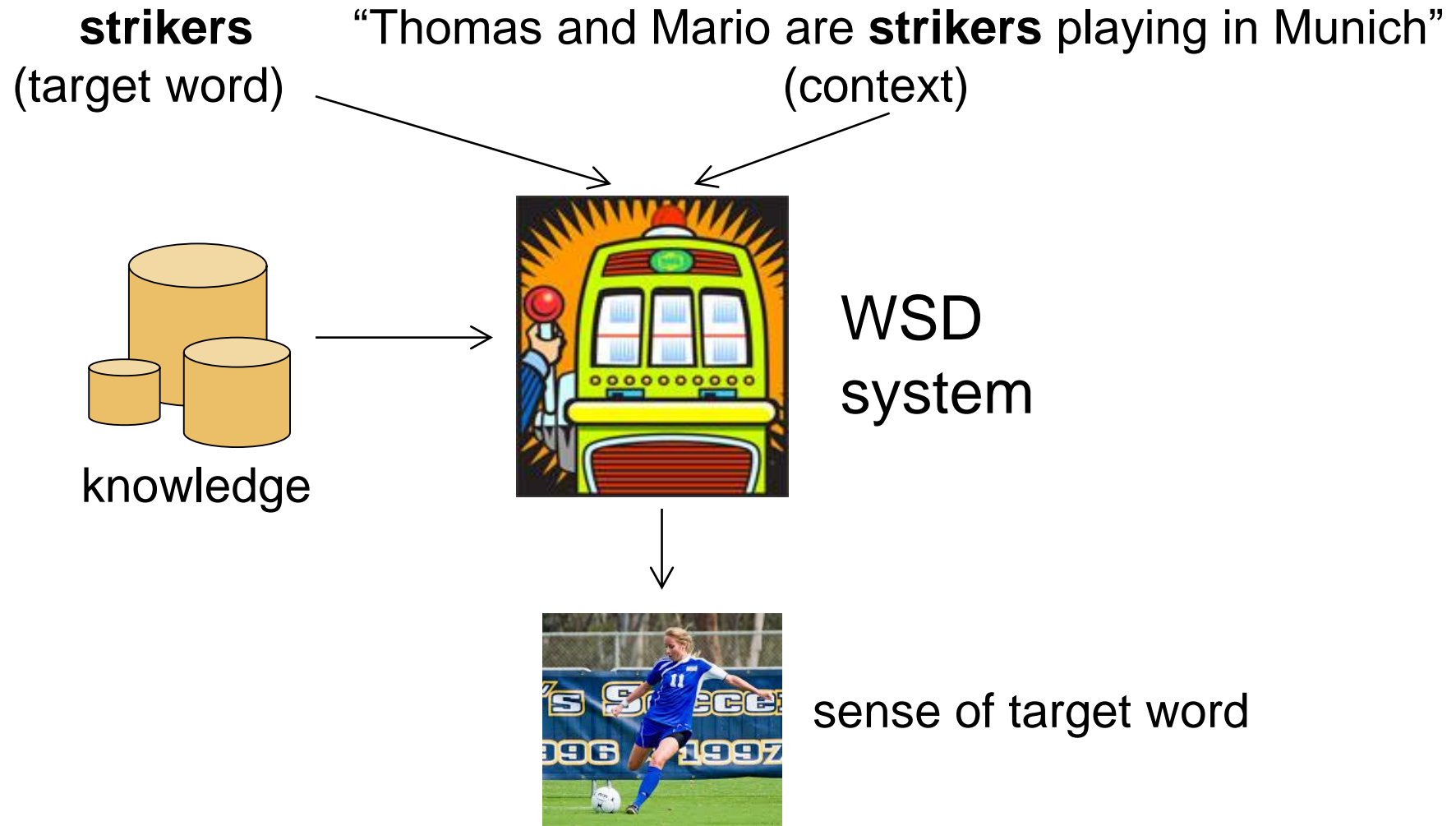
- Thomas and Mario are strikers playing in Munich



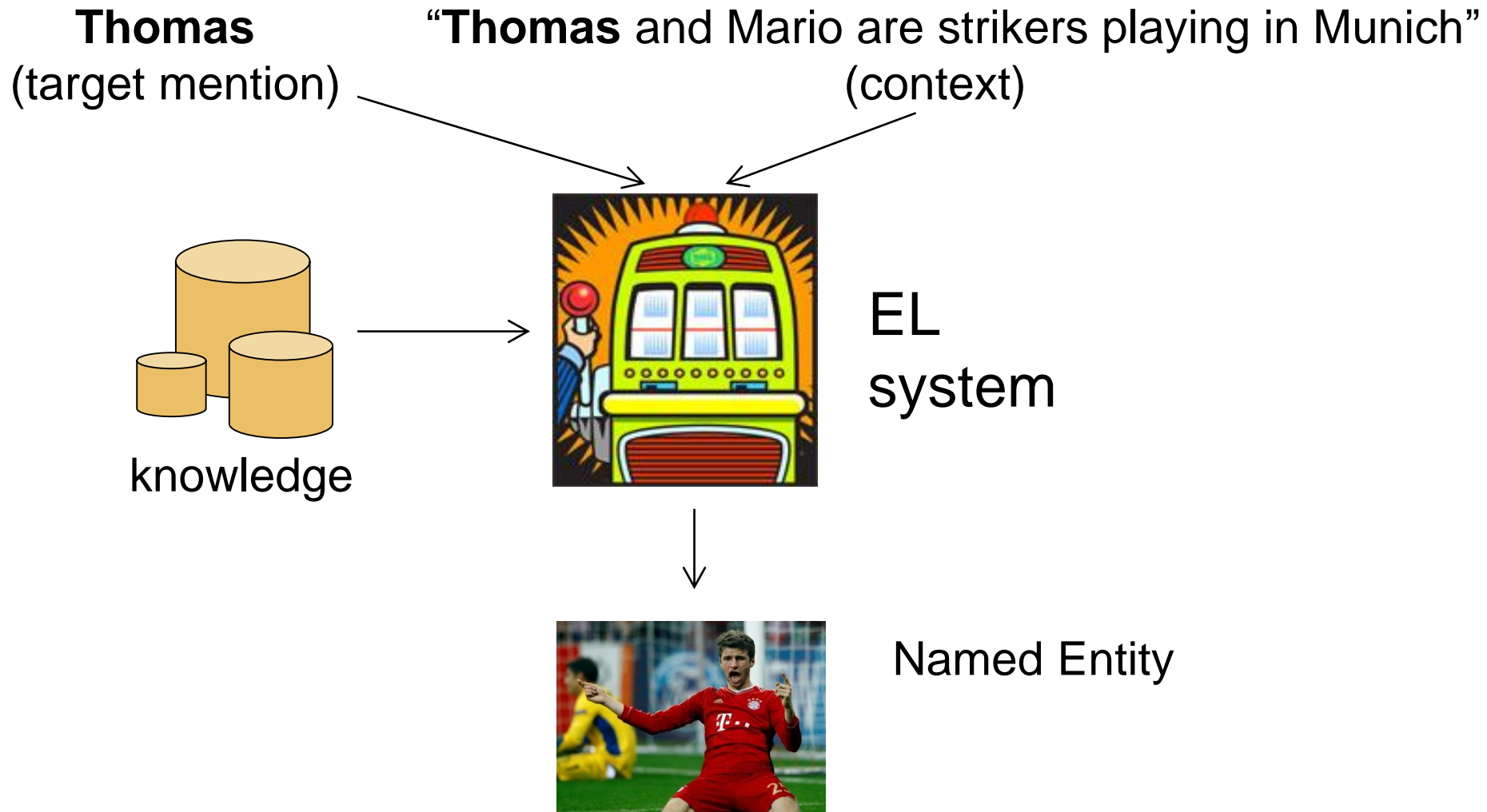
Entity Linking: The task of discovering mentions of entities within a text and linking them in a knowledge base.

WSD: The task aimed at assigning meanings to word occurrences within text.

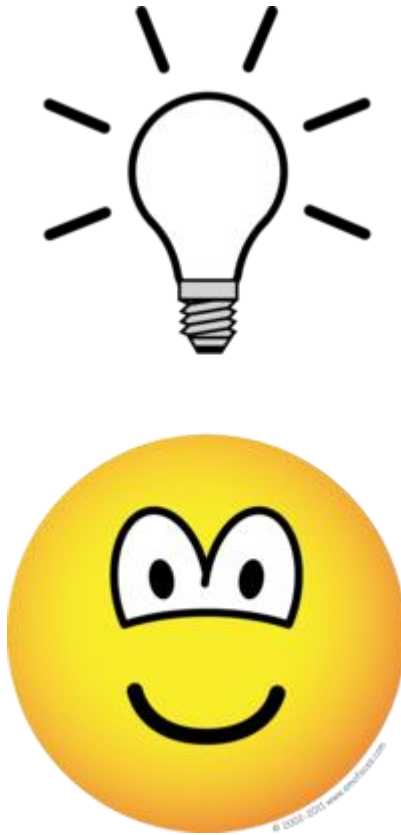
Word Sense Disambiguation in a Nutshell



Entity Linking in a Nutshell



Disambiguation and Entity Linking together!



BabelNet is a huge **multilingual inventory**
for both word senses and named entities!

Multilingual Joint Word Sense Disambiguation (MultiJEDI)

Key Objective 2: use **all languages** to disambiguate **one**



So what?



Babelfy

Step 1: Find all possible meanings of words

1. **Exact Matching** (good for WSD, bad for EL)

~~Thomas~~ and Mario are  ~~soccer~~s playing in Munich



Thomas,
Norman



Thomas,
Seth



They both have
Thomas as one of
their lexicalizations

Step 1: Find all possible meanings of words

2. Partial Matching (good for EL)

Thomas and Mario are strikers playing in Munich



Thomas,
Norman



Thomas,
Seth



Thomas
Müller

It has Thomas as a
substring of one of
its lexicalizations

Step 1: Find all possible meanings of words

“Thomas and Mario are strikers playing in Munich”

Seth Thomas



Mario (Character)



striker (Sport)



Munich (City)



Thomas Müller



Mario (Album)



Striker (Video Game)



FC Bayern Munich



Mario Gómez



Striker (Movie)



Munich (Song)



Thomas (novel)



Step 1: Find all possible meanings of words

“Thomas and Mario are strikers playing in Munich”

Seth Thomas



Mario (Character)



striker (Sport)



Munich (City)



Mario (Album)



Striker (Video Game)



FC Bayern Munich



Thomas Müller



Mario Gómez



Striker (Movie)



Munich (Song)



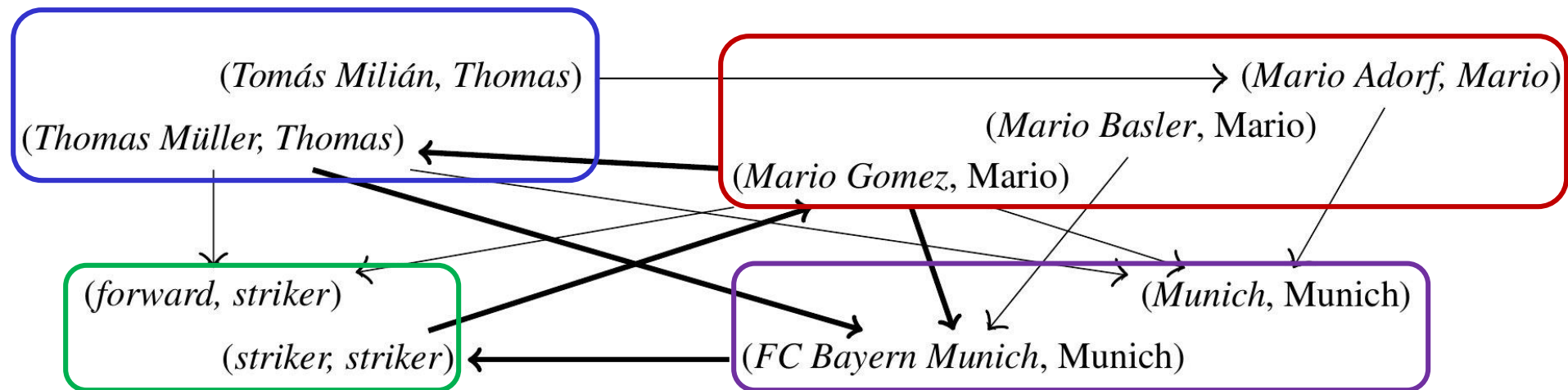
Thomas (novel)



Ambiguity!

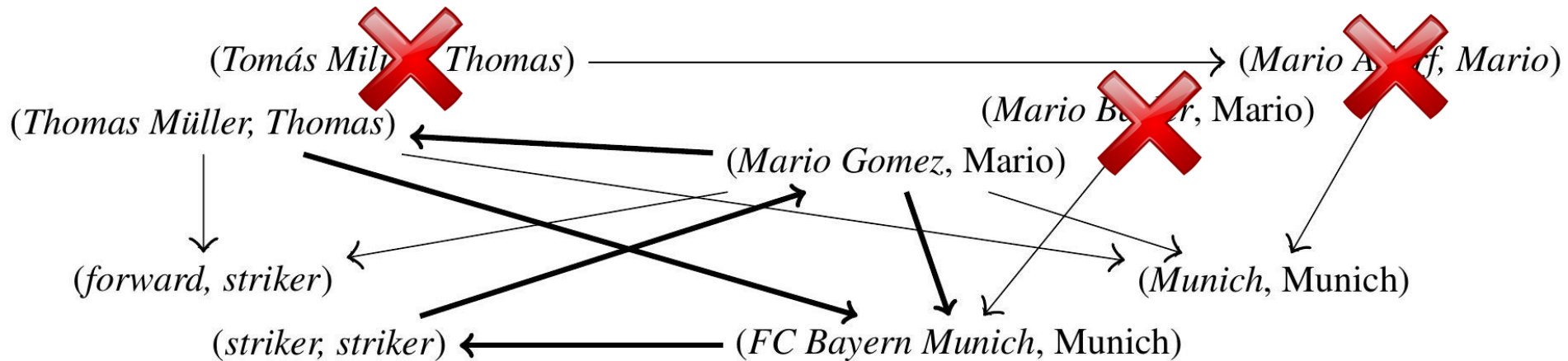
Step 2: Connect all the candidate meanings

Thomas and Mario are strikers playing in Munich



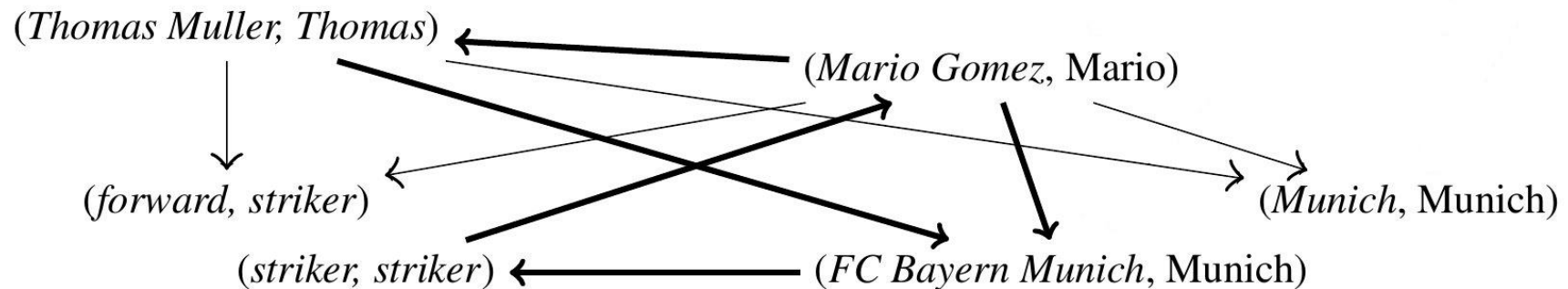
Step 3: Extract a dense subgraph

Thomas and **Mario** are **strikers** playing in **Munich**



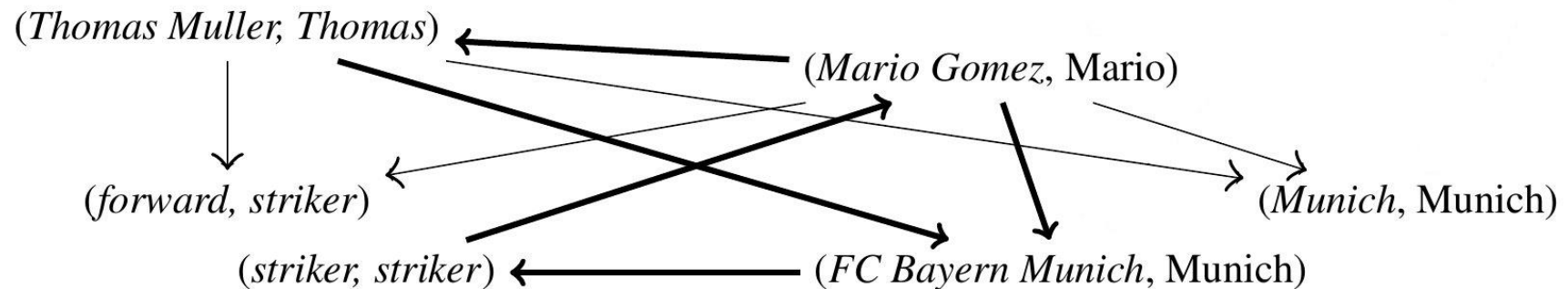
Step 3: Extract a dense subgraph

Thomas and **Mario** are **strikers** playing in **Munich**



Step 4: Select the most reliable meanings

Thomas and **Mario** are **strikers** playing in **Munich**



Step 4: Select the most reliable meanings

“Thomas and Mario are strikers playing in Munich”

Seth Thomas



Mario (Character)



striker (Sport)



Munich (City)



Thomas Müller



Mario (Album)



Striker (Video Game)



FC Bayern Munich



Mario Gómez



Striker (Movie)



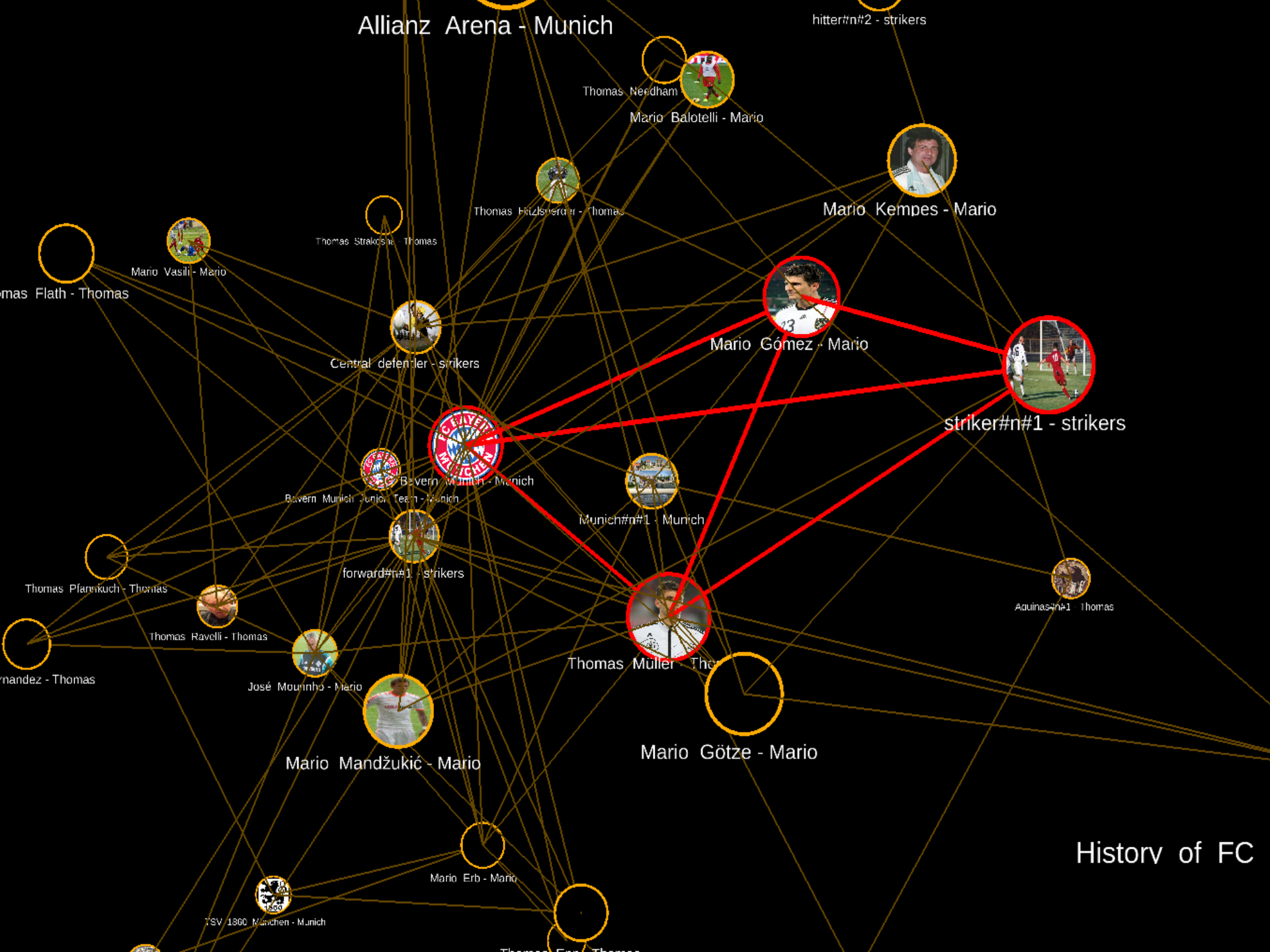
Munich (Song)



Thomas (novel)



Allianz Arena - Munich



History of FC

Experimental Results:

Fine-grained (Multilingual) Disambiguation

Senseval-3 SemEval-2007 task 17 SemEval-2013 task 12

| | Sens3 | Sem07 | SemEval-2013 English | | | French | | German | | Italian | | Spanish | |
|------------------------|-------------|-------------|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| System | WN | WN | WN | Wiki | BN | Wiki | BN | Wiki | BN | Wiki | BN | Wiki | BN |
| Babelfy | 68.3 | 62.7 | 65.9 | 87.4 | 69.2 | 71.6 | *56.9 | 81.6 | 69.4 | 84.3 | 66.6 | 83.8 | 69.5 |
| IMS | 71.2 | 63.3 | 65.7 | – | – | – | – | – | – | – | – | – | – |
| UKB w2w | *65.3 | *56.0 | 61.3 | – | 60.8 | – | 60.8 | – | 66.2 | – | 67.3 | – | 70.0 |
| UMCC-DLSI | – | – | 64.7 | 54.8 | 68.5 | *60.5 | 60.5 | *58.1 | 62.8 | *58.3 | 65.8 | *61.0 | 71.0 |
| DAEBAK! | – | – | – | – | 60.4 | – | 53.8 | – | 59.1 | – | *61.3 | – | 60.0 |
| GETALP-BN | – | – | 51.4 | – | 58.3 | – | 48.3 | – | 52.3 | – | 52.8 | – | 57.8 |
| MFS | 70.3 | 65.8 | *63.0 | *80.3 | *66.5 | 69.4 | 45.3 | 83.1 | *67.4 | 82.3 | 57.5 | 82.4 | *64.4 |
| Babelfy unif. weights | 67.0 | 65.2 | 65.0 | 87.0 | 68.5 | 71.9 | 57.2 | 81.2 | 69.8 | 83.7 | 66.8 | 83.8 | 70.8 |
| Babelfy w/o dens. sub. | 68.3 | 63.3 | 65.4 | 87.3 | 68.7 | 71.6 | 57.0 | 81.7 | 69.1 | 84.4 | 66.5 | 83.9 | 69.5 |
| Babelfy only concepts | 68.2 | 62.7 | 65.5 | 83.0 | 68.7 | 70.2 | 56.6 | 79.3 | 69.3 | 83.0 | 66.3 | 84.0 | 69.7 |
| Babelfy on sentences | 66.0 | 65.2 | 63.5 | 84.0 | 67.1 | 70.7 | 53.6 | 82.3 | 68.1 | 83.8 | 64.2 | 83.5 | 68.7 |

Experimental Results: KORE50, AIDA-CoNLL

- Two gold-standard Entity Linking datasets:

| System | KORE50 | CoNLL |
|------------------------|-------------|-------------|
| Babelfy | 71.5 | 82.1 |
| KORE-LSH-G | 64.6 | 81.8 |
| KORE | 63.9 | *80.7 |
| MW | *57.6 | 82.3 |
| Tagme | 56.3 | 70.1 |
| KPCS | 55.6 | 82.2 |
| KORE-LSH-F | 53.2 | 81.2 |
| UKB w2w (on BabelNet) | 52.1 | 71.8 |
| Illinois Wikifier | 41.7 | 72.4 |
| DBpedia Spotlight | 35.4 | 34.0 |
| Babelfy unif. weights | 69.4 | 81.7 |
| Babelfy w/o dens. sub. | 62.5 | 78.1 |
| Babelfy only NE | 68.1 | 78.8 |

What can we do with Babelfy?

- Disambiguate text written in **any** language!

The screenshot shows the Babelfy web interface. At the top left is the BabelNet logo. To its right is a search bar containing the text "Twitter potenzia chat privata, si può comunicare con tutti." Below the search bar are dropdown menus for "ITALIAN" and "TRANSLATE INTO...". To the right of the search bar are links for "LOG IN" and "REGISTER". Below the search bar is a "SEARCH" button. To the right of the search bar is a "PREFERENCES" link. Below the search bar is a light blue box with the text "This looks like a sentence: loading Babelfy! Please wait." Below this box are two links: "expanded view" and "compact view". Below these links is a horizontal bar with the words "Twitter", "potenzia", "chat", "privata", ", si", "può", "comunicare", and "con". Below this bar are five cards. The first card is for "Twitter" and shows the Twitter logo. The second card is for "chat" and shows a chat window. The third card is for "privata" and shows the text "Confined to particular persons or groups or providing privacy". The fourth card is for "può" and shows the text "To have the ability to do something". The fifth card is for "comunicare" and shows a photo of a man speaking. There are navigation arrows on the left and right sides of the cards.

BabelNet

LOG IN REGISTER

Twitter potenzia chat privata , si può comunicare con

SEARCH

PREFERENCES

This looks like a sentence: loading Babelfy! Please wait.

expanded view | compact view

Twitter potenzia chat privata , si può comunicare con

Twitter

Twitter is an online social networking service that enables users to send and

chat

Online chat may refer to any kind of communication over the Internet that offers

privata

Confined to particular persons or groups or providing privacy

può

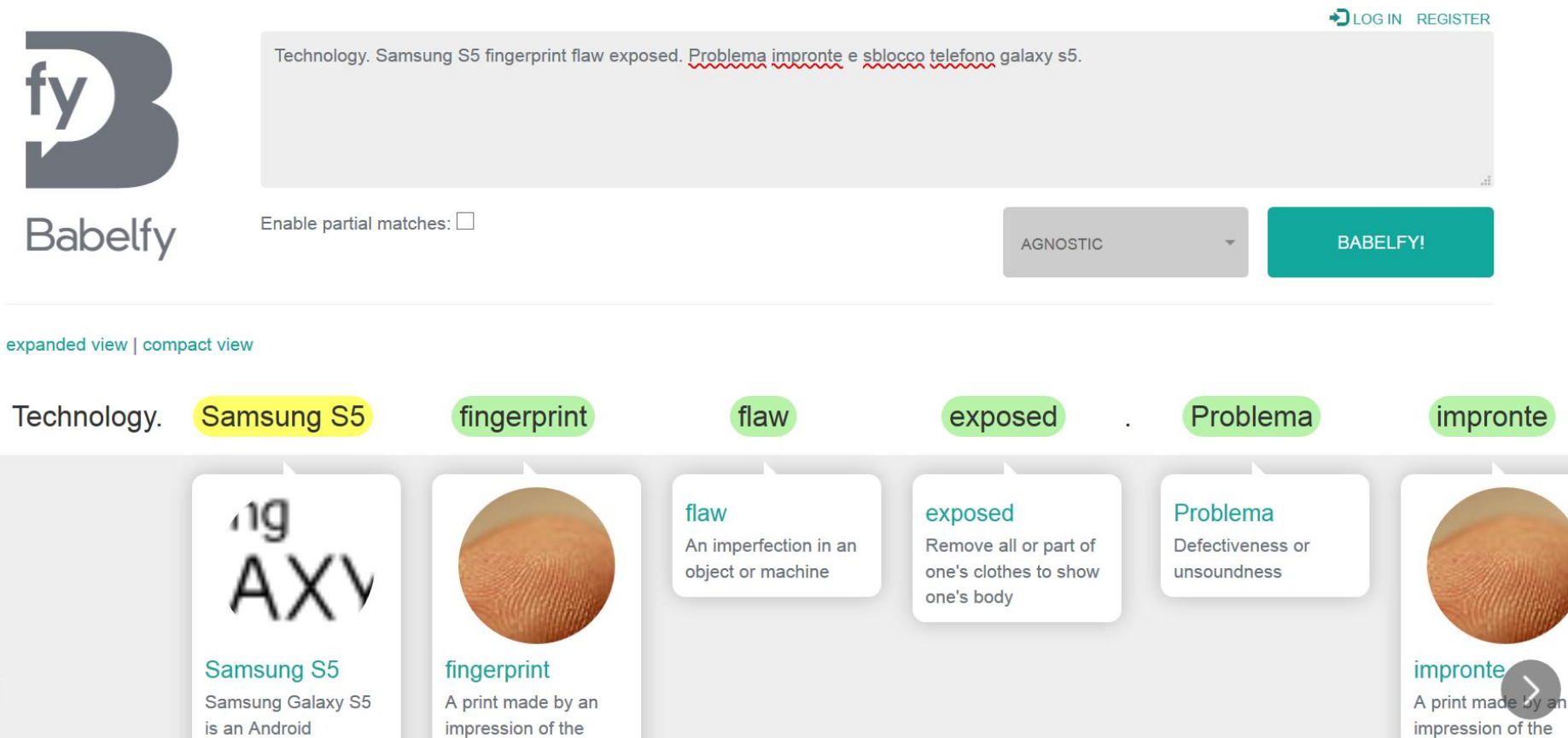
To have the ability to do something

comunicare

Transmit information

What can we do with Babelfy?

- Disambiguate text written in **any** language!
- Disambiguate in a **language-agnostic** setting!



Technology. Samsung S5 fingerprint flaw exposed. Problema impronte e sblocco telefono galaxy s5.

Enable partial matches: ☐

AGNOSTIC

BABELFY!

expanded view | compact view

Technology. Samsung S5 fingerprint flaw exposed . Problema impronte

Samsung S5
Samsung Galaxy S5 is an Android

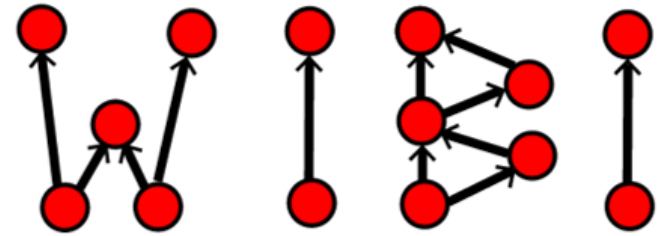
fingerprint
A print made by an impression of the

flaw
An imperfection in an object or machine

exposed
Remove all or part of one's clothes to show one's body

Problema
Defectiveness or unsoundness

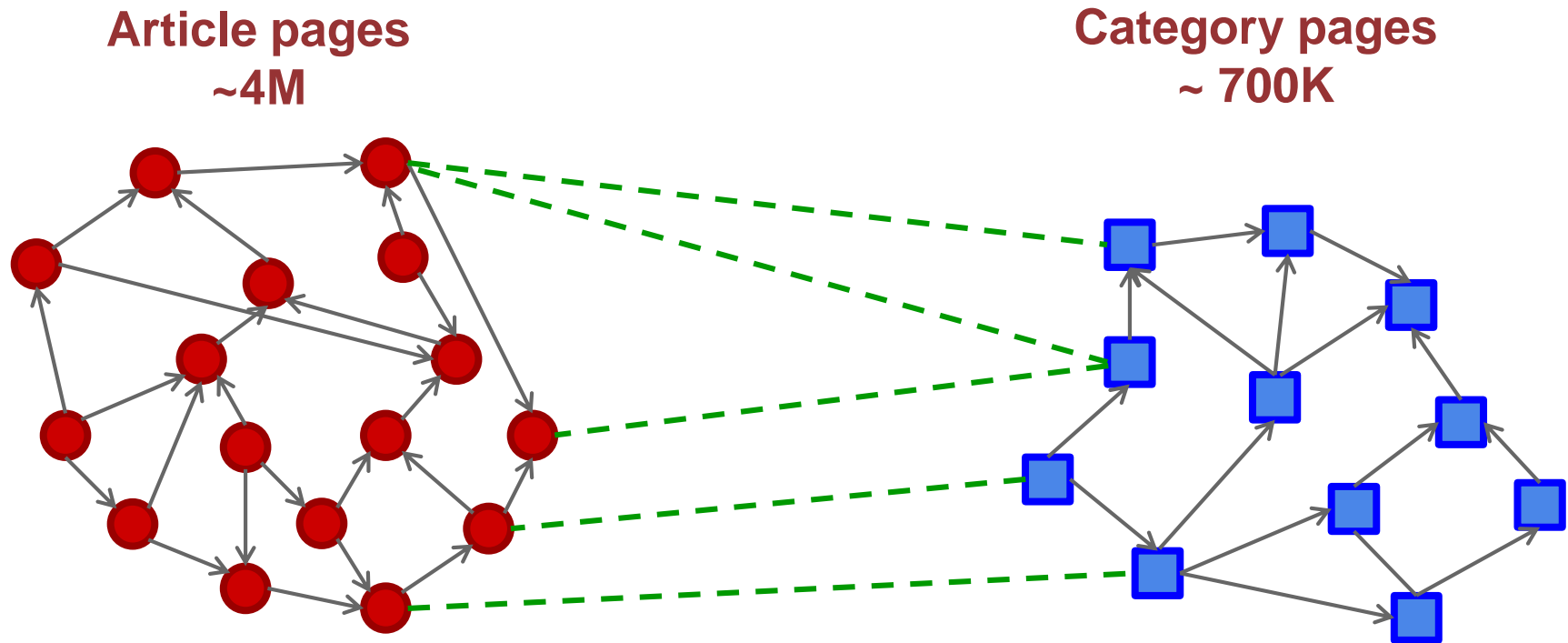
impronte
A print made by an impression of the



STRUCTURING KNOWLEDGE

[Flati et al., ACL 2014]

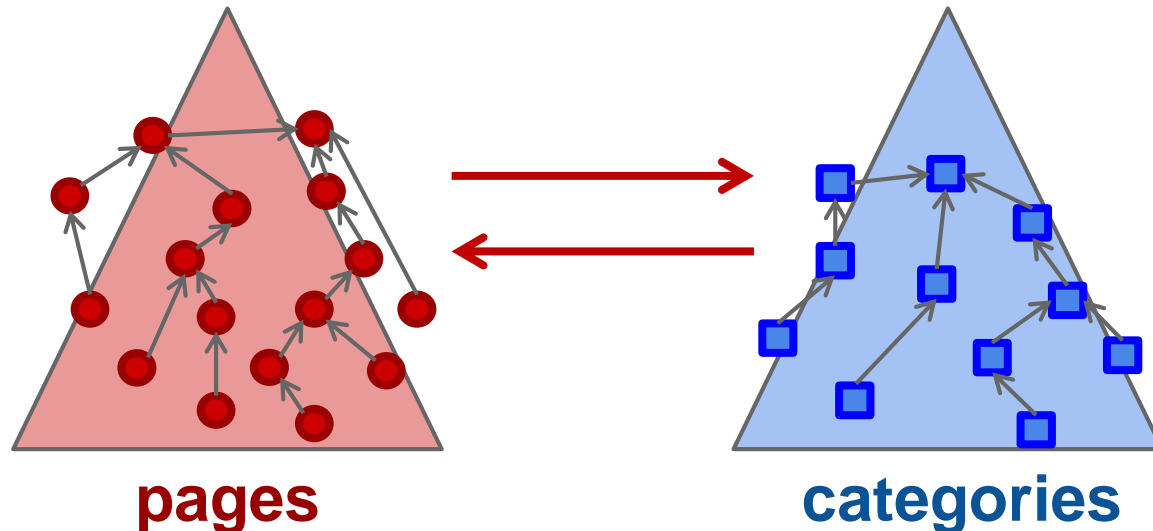
The Wikipedia structure



Two **noisy** graphs with **no** explicit **hypernym** relation.

Our goal

To **automatically** create a **Wikipedia Bitaxonomy** for Wikipedia **pages** and **categories** in a simultaneous fashion.



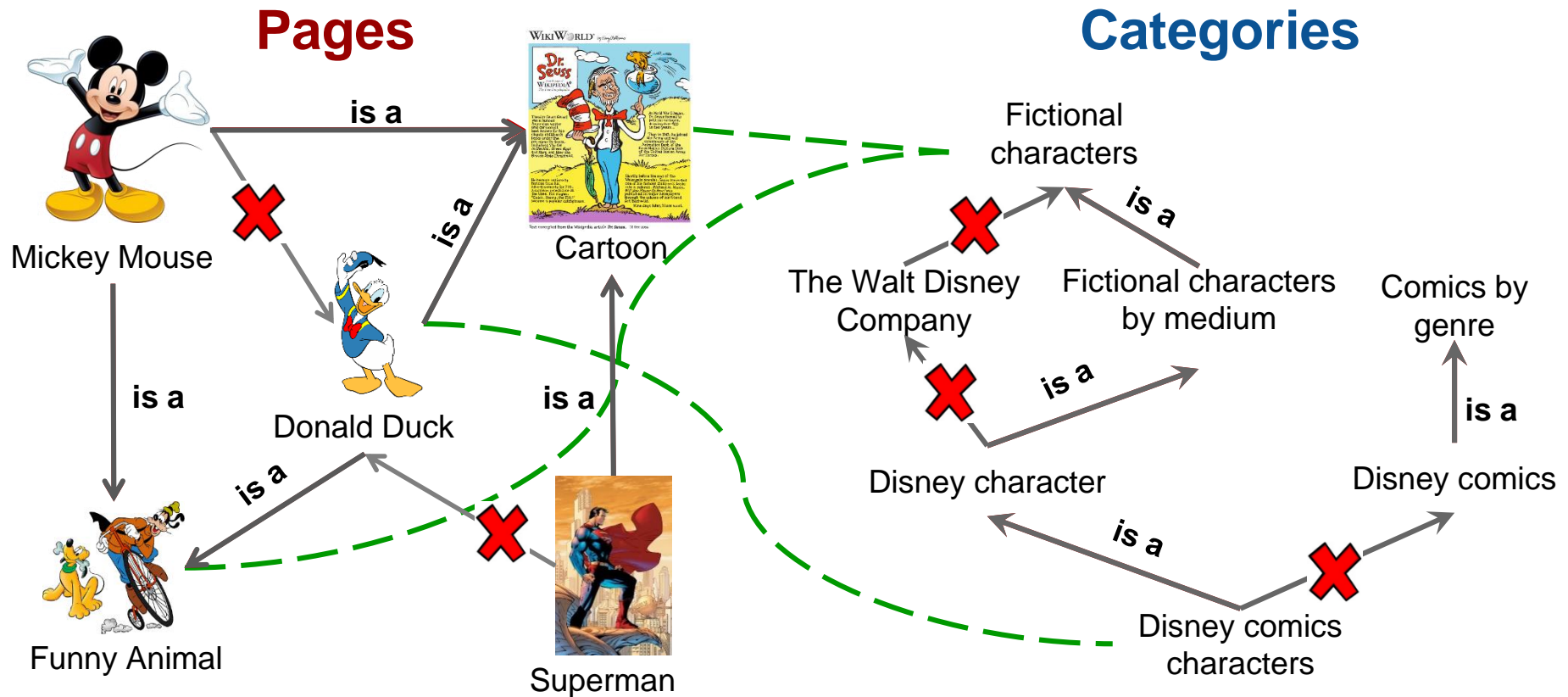
Our goal

To **automatically** create a **Wikipedia Bitaxonomy** for Wikipedia **pages** and **categories** in a simultaneous fashion.

KEY IDEA

The **page** and **category** level are **mutually beneficial** for inducing a **wide-coverage** and **fine-grained** integrated taxonomy

The Wikipedia Bitaxonomy: an example



The Bitaxonomy algorithm

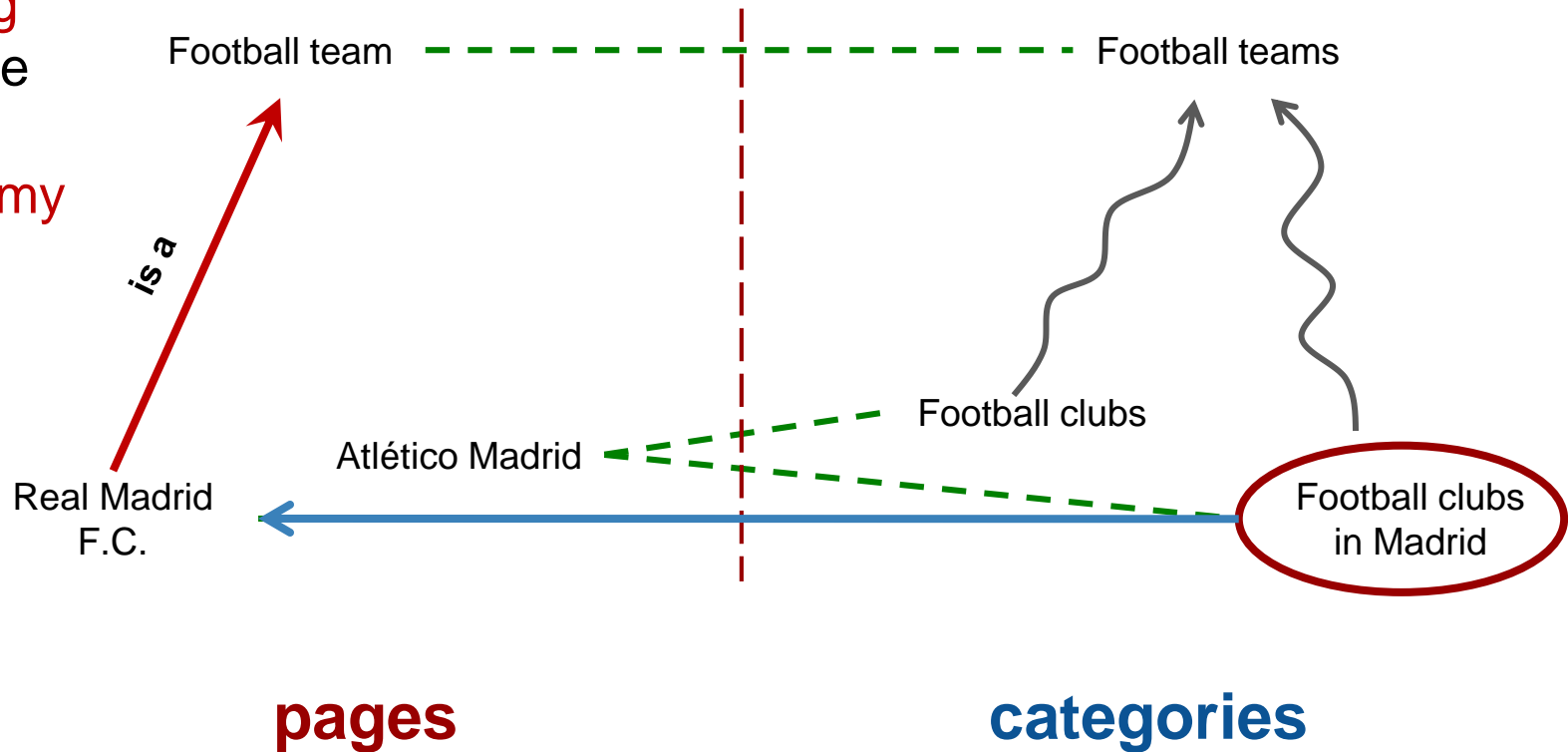
The information available in the two taxonomies is **mutually beneficial**

- At each step **exploit one taxonomy to update the other** and vice versa
- Repeat until **convergence**



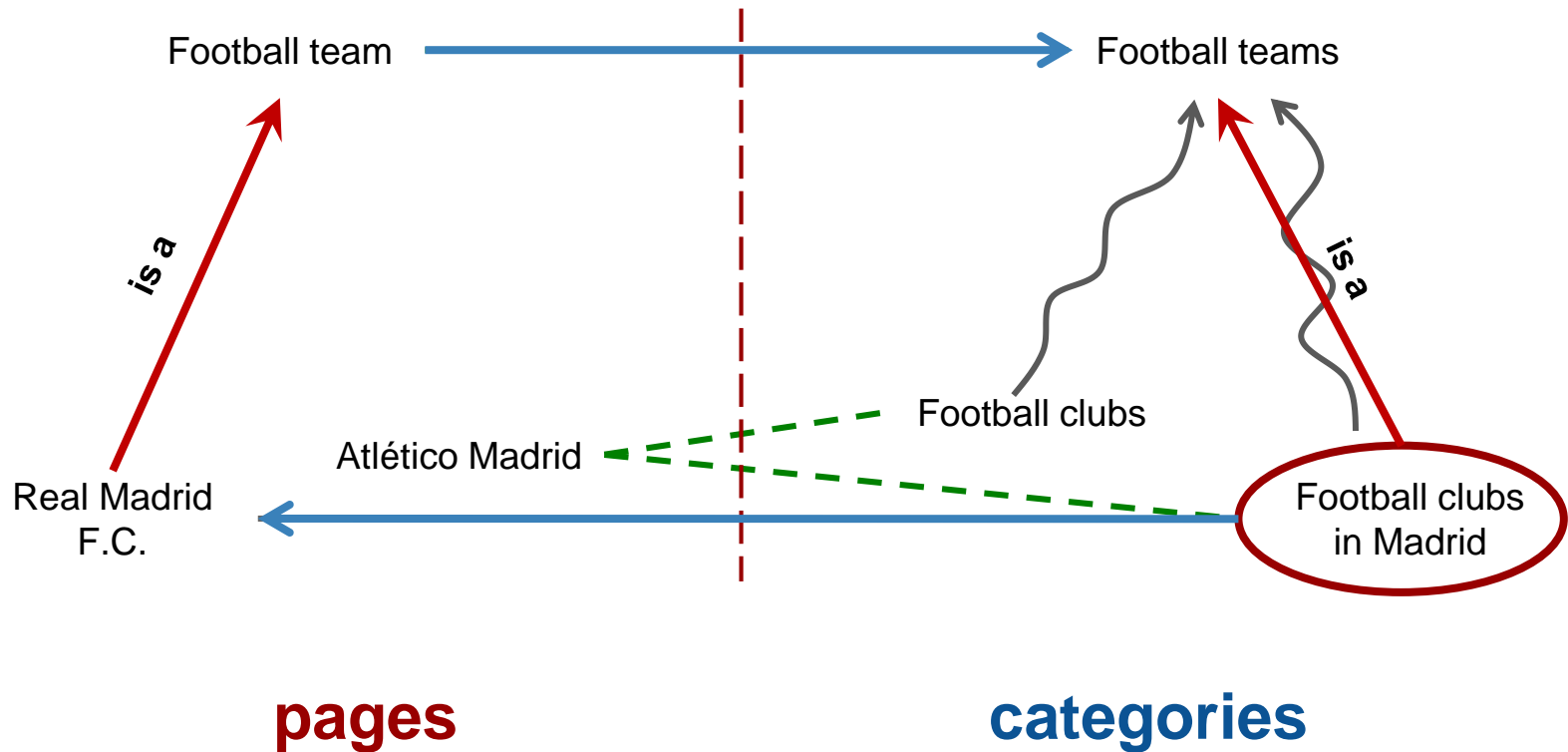
The Bitaxonomy algorithm

Starting
from the
page
taxonomy

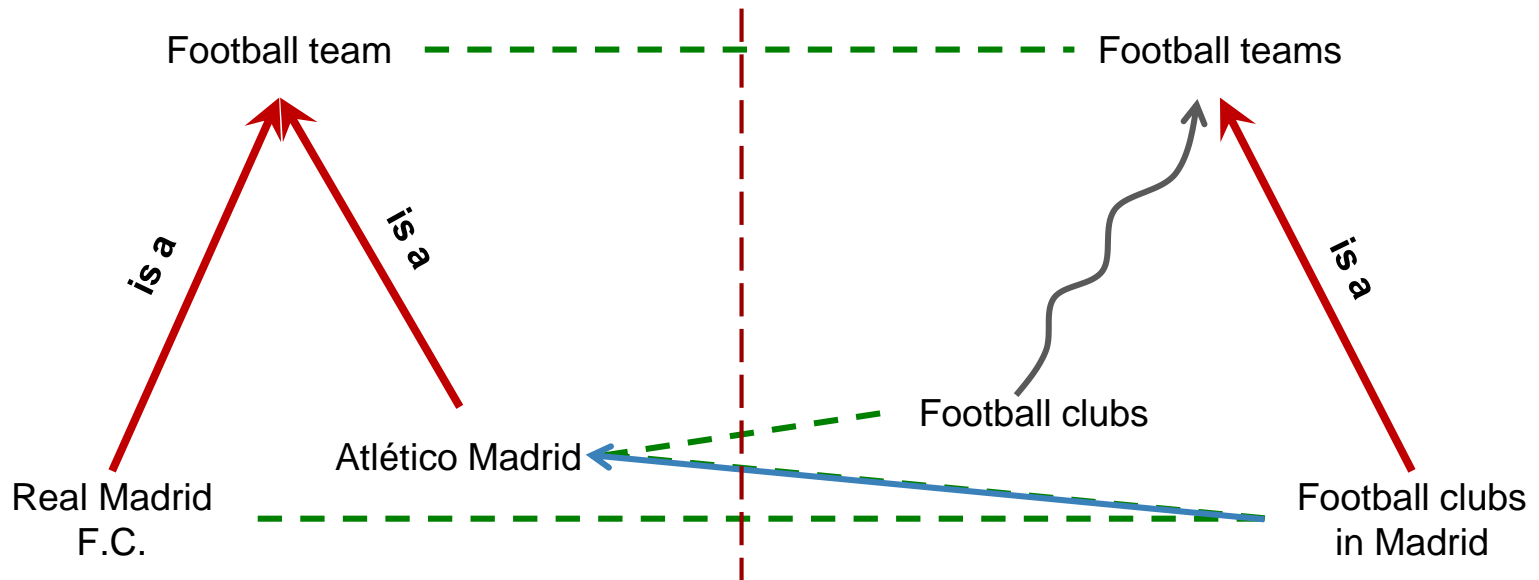


The Bitaxonomy algorithm

Exploit the **cross links** to infer **hypernym** relations in the category taxonomy



The Bitaxonomy algorithm

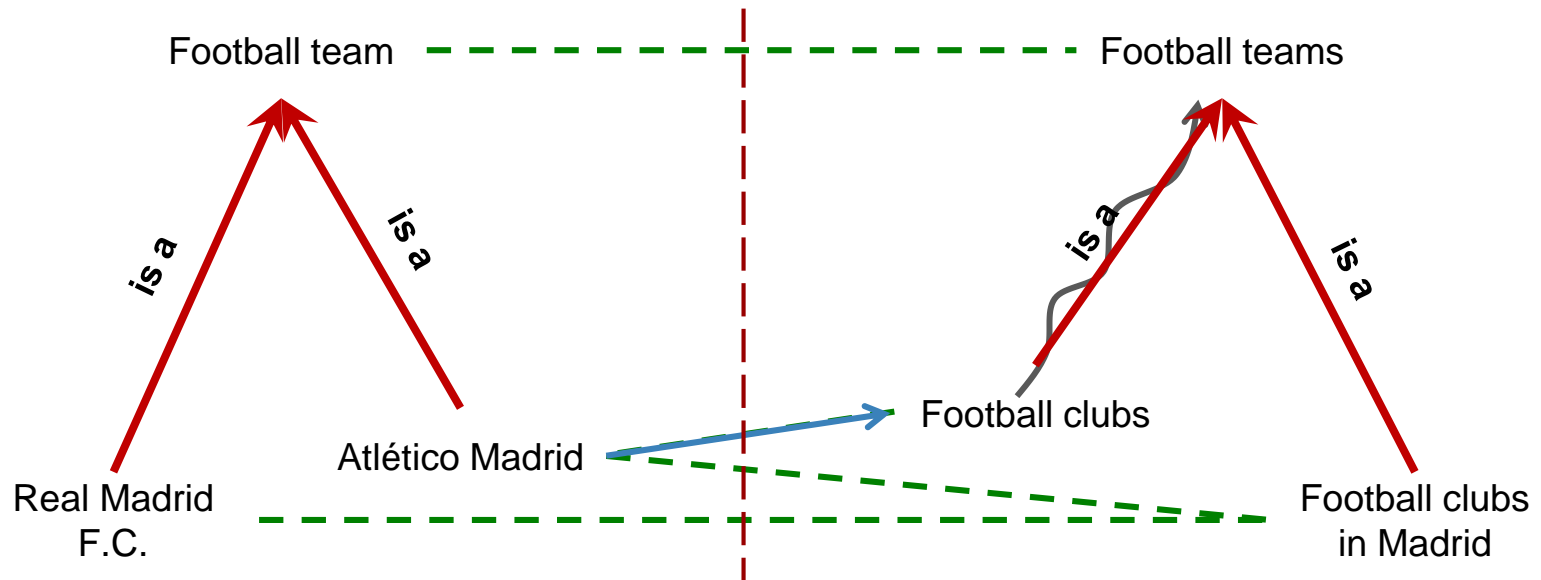


Take advantage of **cross links** to **infer back is-a relations** in the page taxonomy

pages

categories

The Bitaxonomy algorithm



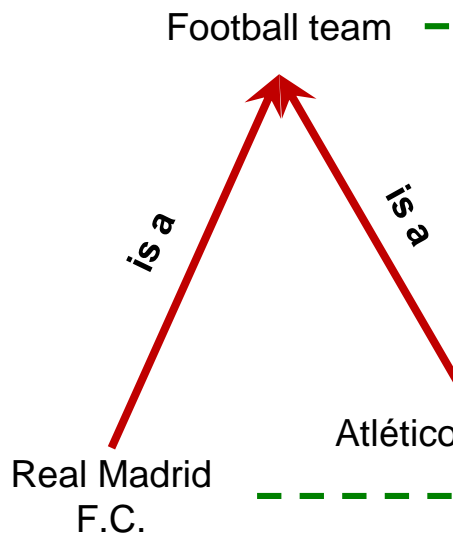
Use the relations found in previous step to infer new hypernym edges

pages

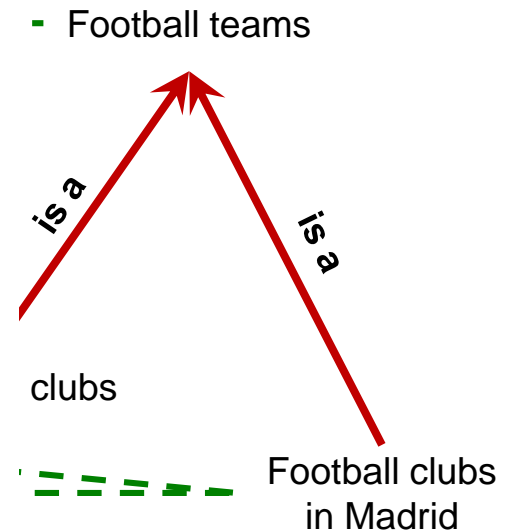
categories

The Bitaxonomy algorithm

Mutual enrichment of both taxonomies until convergence

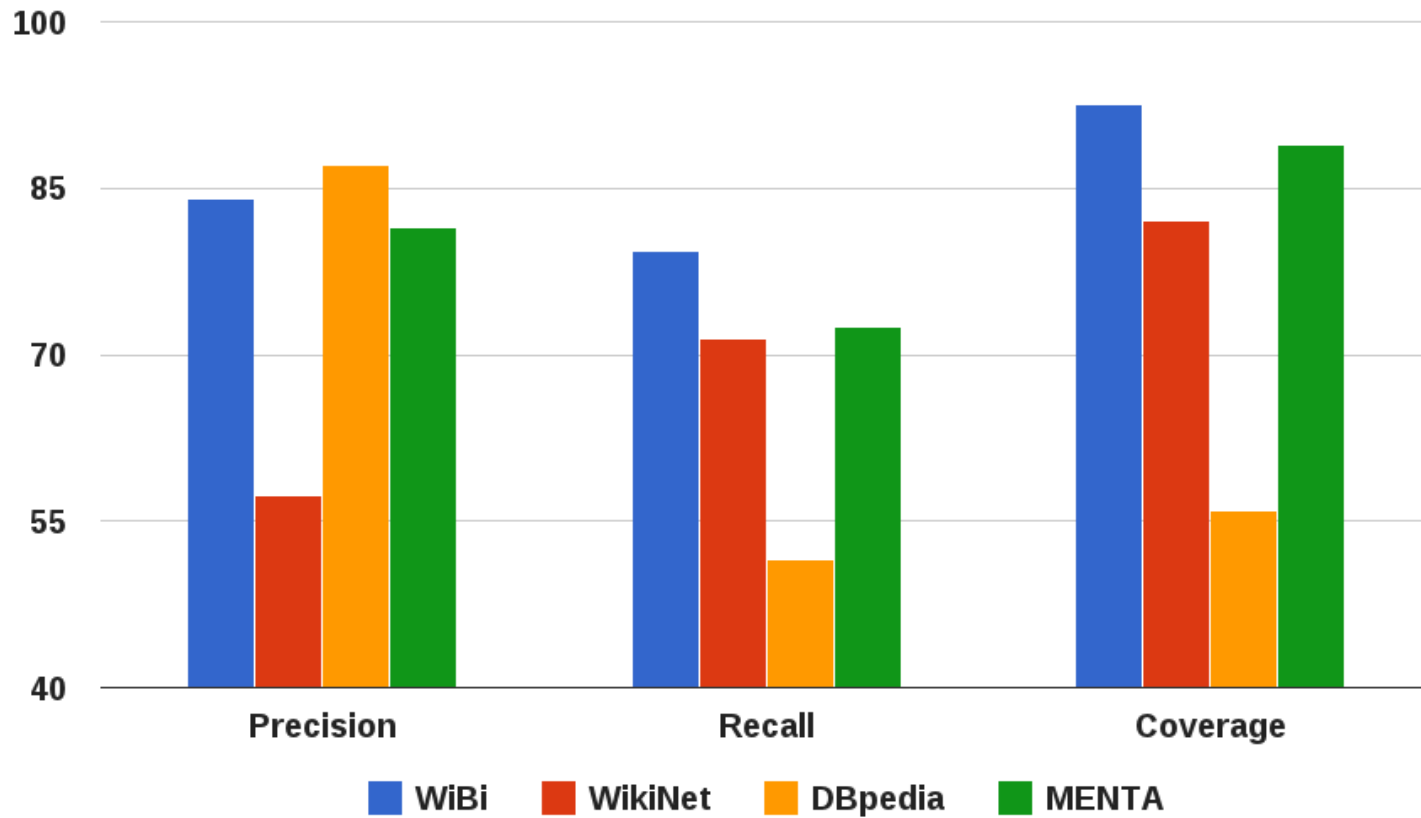


pages

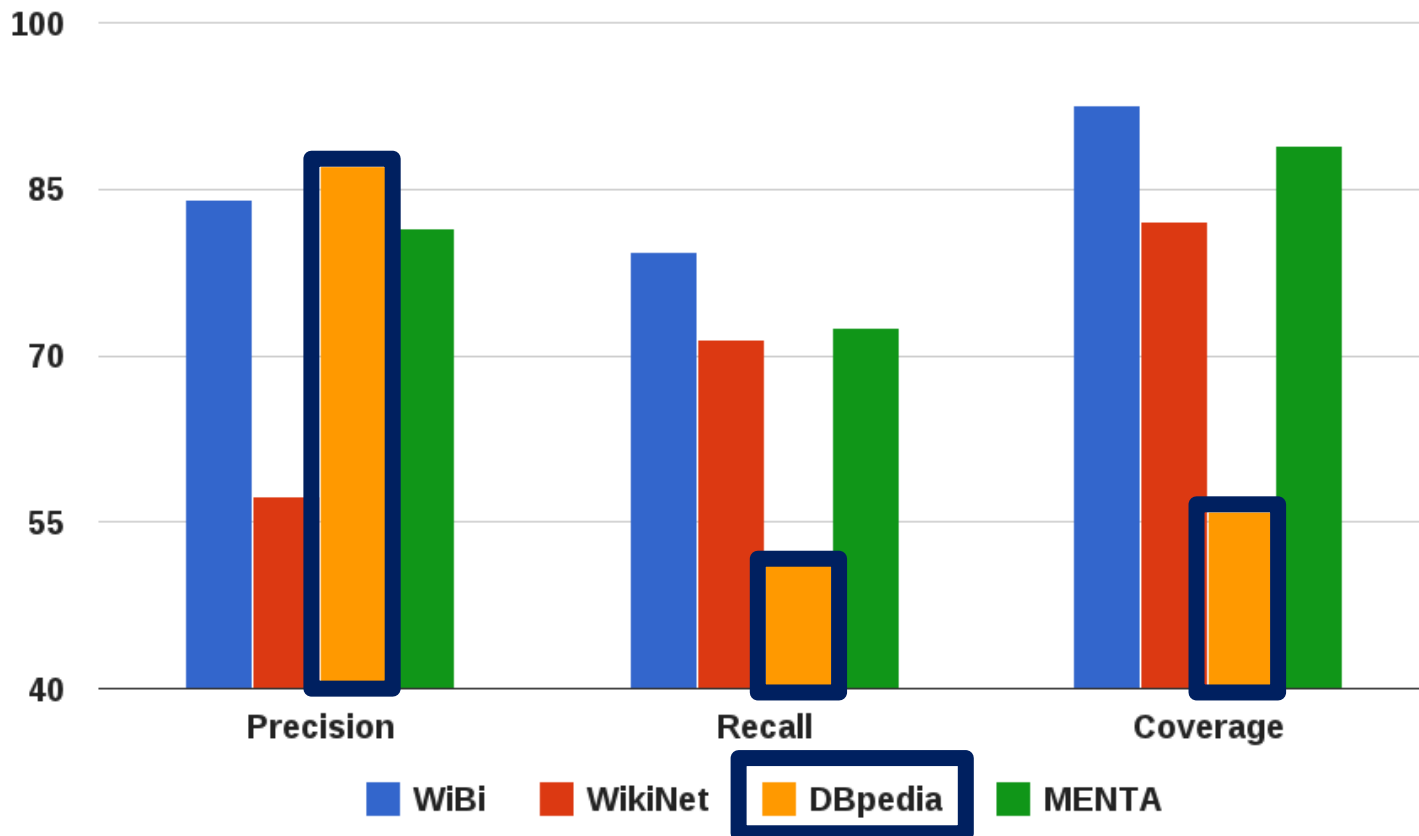


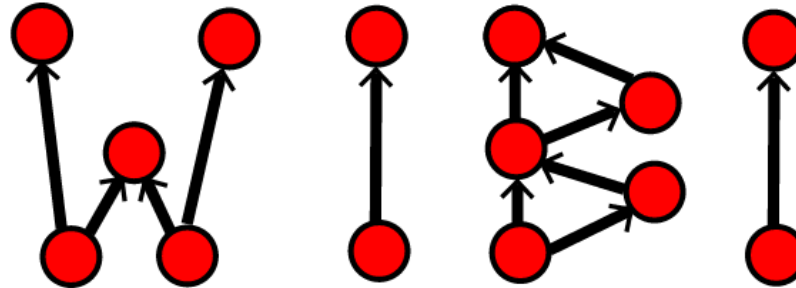
categories

Page Taxonomy Comparison



Page Taxonomy Comparison





WiBi (Wikipedia Bitaxonomy) is an approach to the automatic creation of a bitaxonomy for Wikipedia developed by **Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli.**

WiBi is now also integrated into BabelNet

Input a Wikipedia item

Search

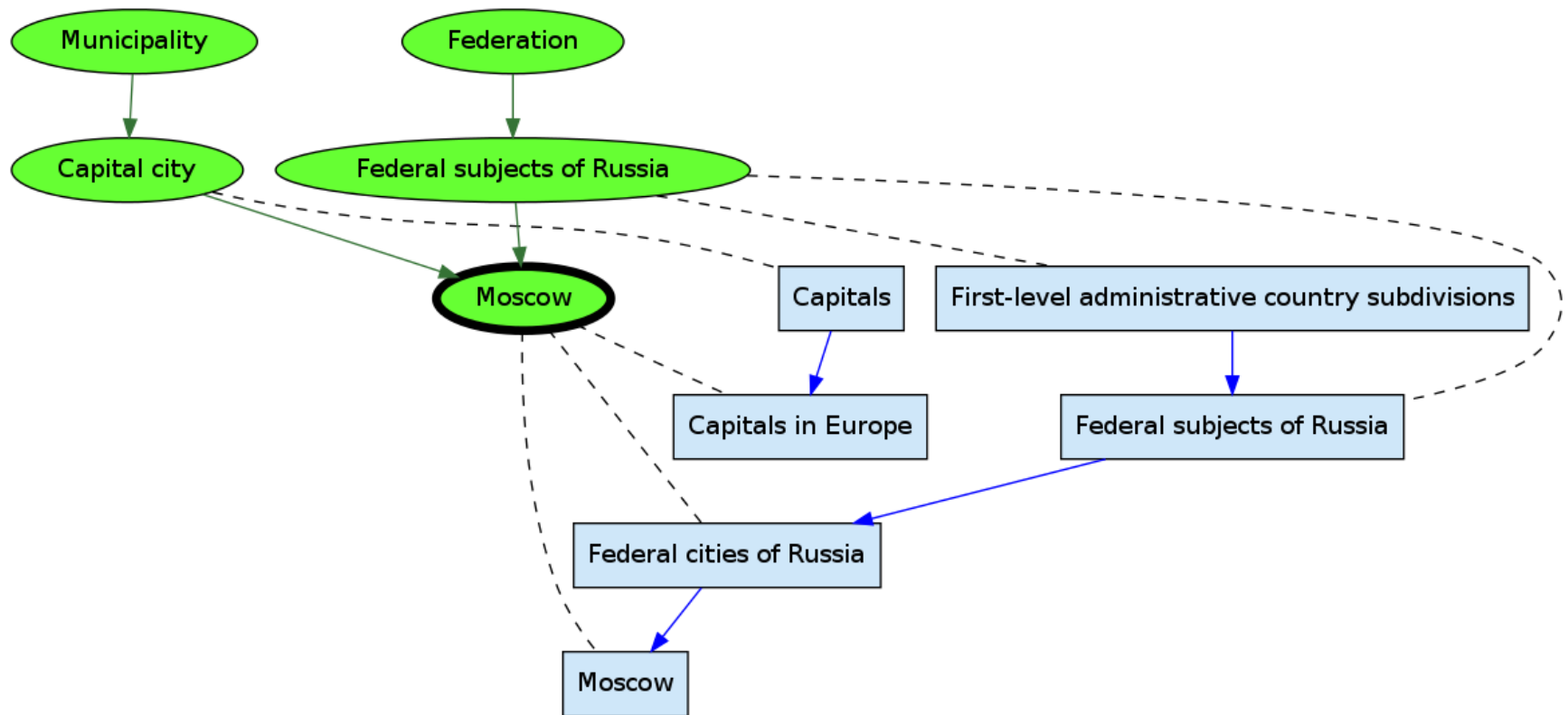
Try out some examples:

[The Da Vinci Code \(film\)](#), [Zucchero Fornaciari](#), [Różyński Wielki](#), [Moulin Rouge](#),
[WordNet](#), [Julia Roberts](#), [Florence](#), [ABBA](#), [Eric Nies](#), [Mąkosy Stare](#)



wibitaxonomy.org


Moscow in the Wikipedia Bitaxonomy



- Also integrated into **BabelNet 3.0**:

bn:00015634n • NOUN • Named Entity • Categories: Moscow, Moscow Governorate, 1147 establishments, Capitals in Europe...

 **Moscow**  • Russian capital • capital of the Russian Federation

A city of central European Russia; formerly capital of both the Soviet Union and Soviet Russia; since 1991 the capital of the Russian Federation  *More definitions*

IS-A: national capital • Federal subjects of Russia • provincial capital • capital PART-OF: Russia

Sense embeddings: explicit meaning and Neural Networks together!!!

- Sense embeddings: **SensEmbed** paper at ACL 2015 [Iacobacci et al. 2015]!
- Disambiguate the **entire English Wikipedia** with Babelfy
- **CBOW**, 5-word window, 400 dimensions, learn **sense embeddings**
- Closest senses to different senses of ambiguous words:

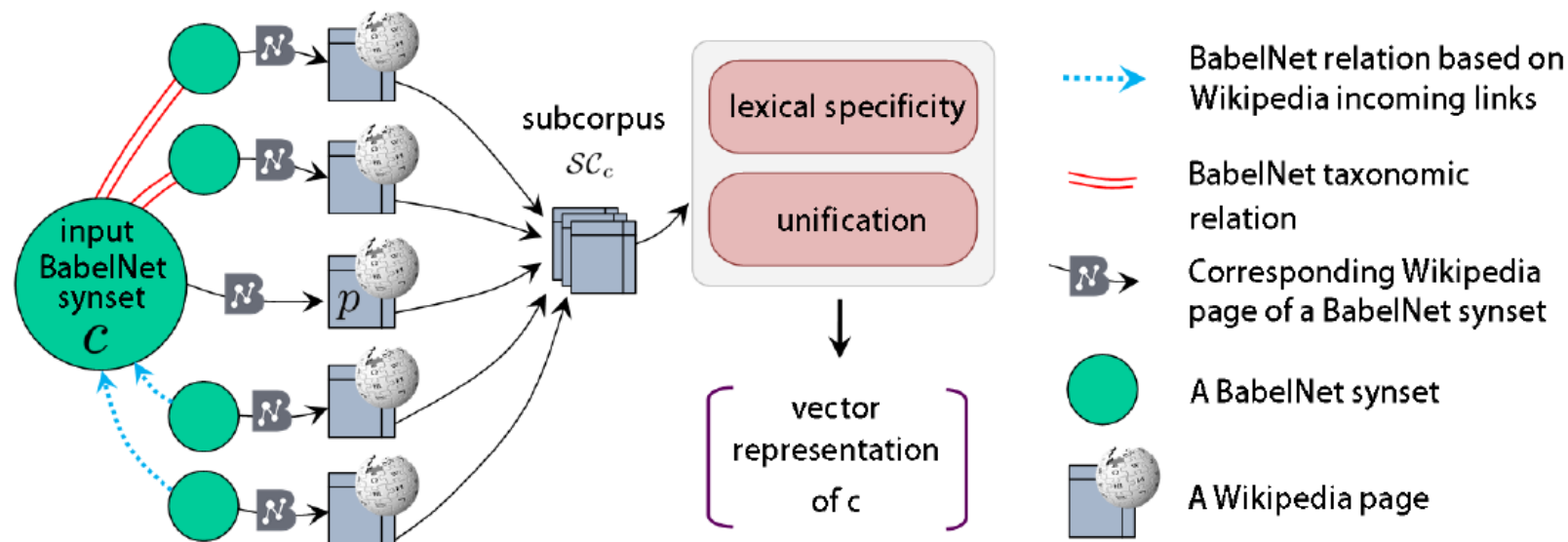
| $bank_1^n$ (geographical) | $bank_2^n$ (financial) | $number_4^n$ (phone) | $number_3^n$ (acting) | $hood_1^n$ (gang) | $hood_{12}^n$ (convertible car) |
|------------------------------|------------------------------|--------------------------|--------------------------|----------------------|------------------------------------|
| upstream $_1^r$ | commercial_bank $_1^n$ | calls $_1^n$ | appearing $_6^v$ | tortures $_5^n$ | taillights $_1^n$ |
| downstream $_1^r$ | financial_institution $_1^n$ | dialled $_1^v$ | minor_roles $_1^n$ | vengeance $_1^n$ | grille $_2^n$ |
| runs $_6^v$ | national_bank $_1^n$ | operator $_{20}^n$ | stage_production $_1^n$ | badguy $_1^n$ | bumper $_2^n$ |
| confluence $_1^n$ | trust_company $_1^n$ | telephone_network $_1^n$ | supporting_roles $_1^n$ | brutal $_1^a$ | fascia $_2^n$ |
| river $_1^n$ | savings_bank $_1^n$ | telephony $_1^n$ | leading_roles $_1^n$ | execution $_1^n$ | rear_window $_1^n$ |
| stream $_1^n$ | banking $_1^n$ | subscriber $_2^n$ | stage_shows $_1^n$ | murders $_1^n$ | headlights $_1^n$ |

Sense embeddings: explicit meaning and Neural Networks together!!!

- Sense embeddings: **SensEmbed** paper at ACL 2015 [Iacobacci et al. 2015]!
- State-of-the-art performance **beyond word2vec**:

| Measure | Dataset | | | | | Average |
|--------------------------------|--------------|--------------|--------------|--------------|--------------|---------|
| | RG-65 | WS-Sim | WS-Rel | YP-130 | MEN | |
| Pilehvar et al. (2013) | 0.868 | 0.677 | 0.457 | 0.710 | 0.690 | |
| Zesch et al. (2008) | 0.820 | — | — | 0.710 | — | |
| Collobert and Weston (2008) | 0.480 | 0.610 | 0.380 | — | 0.570 | |
| Word2vec (Baroni et al., 2014) | 0.840 | 0.800 | 0.700 | — | 0.800 | |
| GloVe | 0.769 | 0.666 | 0.559 | 0.577 | 0.763 | |
| ESA | 0.749 | — | — | — | — | |
| PMI-SVD | 0.738 | 0.659 | 0.523 | 0.337 | 0.726 | |
| Word2vec | 0.732 | 0.707 | 0.476 | 0.343 | 0.665 | |
| SENSEMBED _{closest} | 0.894 | 0.756 | 0.645 | 0.734 | 0.779 | 0.770 |
| SENSEMBED _{weighted} | 0.871 | 0.812 | 0.703 | 0.639 | 0.805 | 0.794 |

MUFFIN: Multilingual UniFied Flexible Interpretation [Camacho-Collados et al., ACL 2015]



- Unification is based on the **Wikipedia Bitaxonomy**
- We obtain an **explicit semantic vector** for each BabelNet synset (**multilingual** and **unified!**)

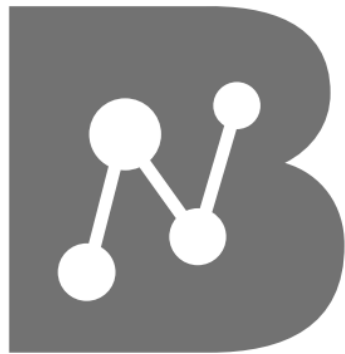
MUFFIN: Multilingual UniFied Flexible Interpretation [Camacho-Collados et al., ACL 2015]

- Performs consistently well across languages:

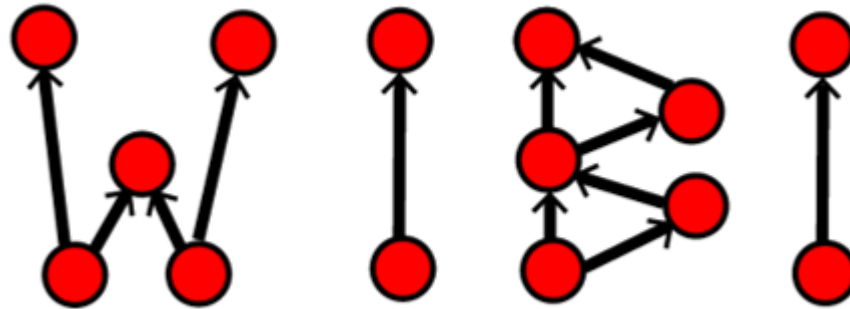
| English | ρ | r | German | ρ | r | French | ρ | r |
|--------------|-------------|-------------|--------------|-------------|-------------|--------------|-------------|-------------|
| MUFFIN | 0.83 | 0.84 | MUFFIN | 0.77 | 0.76 | MUFFIN | 0.71 | 0.77 |
| SOC-PMI | – | 0.61 | SOC-PMI | – | 0.27 | SOC-PMI | – | 0.19 |
| PMI | – | 0.41 | PMI | – | 0.40 | PMI | – | 0.34 |
| Retrofitting | 0.74 | – | Retrofitting | 0.60 | – | Retrofitting | 0.61 | – |
| LSA-Wiki | 0.69 | 0.65 | – | – | – | LSA-Wiki | 0.52 | 0.57 |
| Wiki-wup | – | 0.59 | Wiki-wup | – | 0.65 | | | |
| SSA | 0.83 | 0.86 | Resnik | – | 0.72 | | | |
| NASARI | 0.84 | 0.82 | Lesk_hyper | – | 0.69 | | | |
| ADW | 0.87 | 0.81 | | | | | | |
| Word2Vec | – | 0.84 | | | | | | |
| PMI-SVD | – | 0.74 | | | | | | |
| ESA | – | 0.72 | | | | | | |

Spearman (ρ) and Pearson (r) correlation performance of different systems on the English, German and French RG-65 datasets.

Summarizing



BabelNet



+ preview on **sense embeddings** and **explicit multilingual vectors** for state-of-the-art semantic similarity!

NOW...

**WHAT WILL YOU DO WITH
BABELNET TONIGHT?**

Thanks or...





SAPIENZA
UNIVERSITÀ DI ROMA

Roberto Navigli

Linguistic Computing Laboratory
<http://lcl.uniroma1.it>

