Babelfying the Multilingual Web: state-of-the-art Disambiguation and Entity Linking of Web pages in 50 languages

Roberto Navigli
navigli@di.uniroma1.it

DIPARTIMENTO
DI INFORMATICA

SAPIENZA
UNIVERSITÀ DI ROMA

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

Joint work with

Andrea Moro

Alessandro Raganato

A. Moro, A. Raganato, R. Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. **Transactions of the Association for Computational Linguistics (TACL)**, 2, 2014.

I WANT TO READ THE MEME
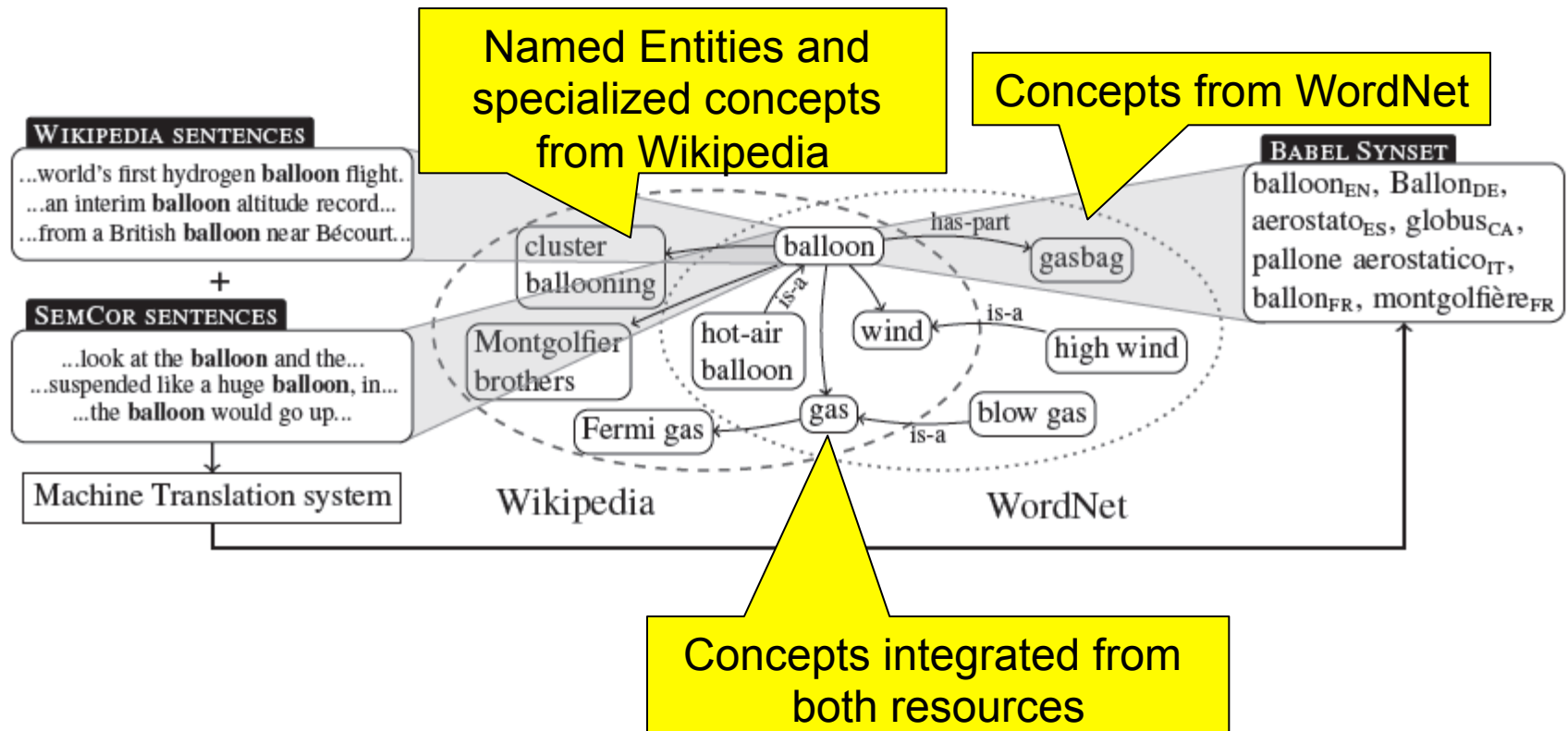BUT IT'S WRITTEN IN TRADITIONAL CHINESE

Motivation

- Domain-specific Web content is available in **many languages**

- Information should be extracted and processed **independently** of the source/target language

- This *could* be done automatically by means of high-performance **multilingual text understanding**

BabelNet (http://babelnet.org)

A wide-coverage multilingual semantic network and encyclopedic dictionary in 50 languages!

BabelNet (http://babelnet.org)

New 2.5 version out!

- **50** languages covered (including Latin!)

- **21 million** textual definitions

- **67M** word senses and named entities!

- **1.1B** RDF triples available via SPARQL endpoint

- Seamless integration of:
    - **WordNet** 3.0
    - **Wikipedia**
    - **Wiktionary**
    - **OmegaWiki**: a collaborative multilingual dictionary
    - **Open Multilingual WordNet** [Bond and Foster, 2013]

- Translations for all open-class parts of speech

BabelNet as a Multilingual Inventory for

Concepts

Calcio in Italian can denote different concepts:
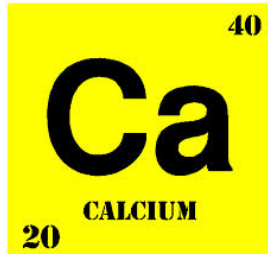
Named Entities

The text *mario* can be used to represent different things such as the video game charachter or a soccer player (Gomez) or even a music album

Calcio / Kick in BabelNet 2.5

Meaning: kick[1] • **ID:** bn:00012171n • **Type:** Concept [details] [explore]

Senses:

🇬🇧 kick[1], boot[7], kicking[2]

🇵🇸 رَفْس، رَفَسَ، رَكْل، 🇫🇷 coup de pied, 🇬🇷 κλοτσιά, λάκτισμα, 🇮🇹 **calcio**, pedata, scalciata, 🇯🇵 キック, 一蹴り, 足蹴, 蹴り

🇬🇧 Kick, 🇨🇳 脚法, 🇫🇷 Coup de pied, 🇩🇪 Tritttechnik, 🇮🇱 בעיטה, 🇮🇹 **Calcio (pedata)**, 🇯🇵 蹴り技, 🇷🇺 Удар ногой, 🇪🇸 Patada

🇬🇧 Kicks, Back Kick, Kicking, Push Kick, Hook Kick, 🇨🇳 腳法, 踢 (武術), 足技, 🇫🇷 Coups de pied, Coup De Pied, 🇯🇵 かかと落とし, キック, 踵落とし, 蹴り, 蹴る, ネリチャギ, 🇪🇸 Patadas, Patear

['w] 🇬🇧 kick, boot, 🇮🇱 בעיטה, 🇷🇺 пинок

🇬🇧 kick, 🇫🇷 Coup de pied, 🇩🇪 Tritttechnik, 🇮🇱 בעיטה, 🇮🇹 **calcio**, 🇯🇵 蹴り技, 🇷🇺 Удар ногой, 🇪🇸 Patada

🇨🇳 踢, 🇫🇷 coup de pied, coups de pied, 🇩🇪 kick, treten, 🇬🇷 κλωτσιά, κλωτσιές, κλωτσιές, 🇮🇱 בעיטה, 🇮🇳 लात, kicks, लात मार, 🇮🇹 **calcio**, calci, 🇯🇵 キック、, 🇷🇺 удар, ногами, 🇪🇸 patadas

Glosses:

['w] 🇬🇧 A hit or strike with the leg or foot or knee.

['w] 🇬🇧 A blow with the foot; a kick.

🇬🇧 Physical strike in combat sports

🇬🇧 The act of delivering a blow with the foot; "he gave the ball a powerful kick"; "the team's kicking was excellent"

🇬🇧 In combat sports and hand-to-hand combat, a kick is a physical strike using the foot, leg, or knee.

🇫🇷 Un coup de pied en sport de combat est une frappe réalisée avec une action du membre inférieur dont la surface d'impact est le pied, Dans certaines pratiques de combat, on parle de techniques de jambe; elles sont la partie émergente de leur activité, ce qui est le cas pour le taekwondo moderne.

Babelfying the Multilingual Web: state-of-the-art Disambiguation and Entity Linking
Andrea Moro, Alessandro Raganato and **Roberto Navigli**

Calcio / Calcium in BabelNet 2.5

Meaning: calcium[1] · **ID:** bn:00006850n · **Type:** Concept · [details] [explore]

Senses:

🇬🇧 calcium[1], Ca[1], atomic number 20[1]

العنصر العشرون, الكّلسيوُم, كاّلسيوُم , كّلسيوُم, 🇫🇷 calcium, 🇬🇷 ασβέστιο, 🇮🇱 סידן, 🇮🇹 **calcio**, 🇯🇵 Ca, カルシウム, Ca

🇬🇧 Calcium, Calcium in biology, كالسيوم, 🇨🇳 钙, 鈣営養, 🇫🇷 Calcium, 🇩🇪 Calcium, 🇬🇷 Ασβέστιο, 🇮🇱 סידן, 🇮🇳 कैल्शियम, 🇮🇹 **Calcio (elemento)**, 🇯🇵 カルシウム, 🇷🇺 Кальций, 🇪🇸 Calcio

🇬🇧 Calcium storage, Ca(2+), Calcium (biology), Calcium storages, Factor 4, Kalcium, Calium, Calcium ions, Ca2+, Ca storages, Ca (biology), Calcium (second messenger), Ca+2, Ca ion (physiology), Calcium supplements, Calcium compounds, Calcium physiology, Biological role of calcium, Ca2, Ca²⁺, Calcium ion physiology, Element 20, الكالسيوم, Ca, Calcium, 🇨🇳 鈣, 20號元素, 元素20, 第20號元素, Calcium, 鈣質, 🇫🇷 Calcium (nutriment), 7440-70-2, Oxydation du calcium, 🇩🇪 Ca++, Ca2+, Kalzium, 🇮🇱 קאלציום, קצ'ום, 🇮🇳 कैलशियम, कालिसयम, कैल्सियम, Ca, 🇮🇹 **Calcio (elemento chimico)**, **Calcio (metallo)**, Fattore di coagulazione IV, **Calcio (chimica)**, Ione calcio, Fattore coagulante IV, 🇯🇵 カルシウムイオン, カルシューム, 🇷🇺 Ca, Ca2+, 🇪🇸 Factor de coagulación IV, Calcios

🇬🇧 calcium, كالسيوم, 🇫🇷 calcium, 🇩🇪 Kalzium, Calcium, 🇬🇷 ασβέστιο, 🇮🇱 סידן, 🇮🇳 कालिसयम, 🇮🇹 **calcio**, 🇯🇵 カルシウム, 🇷🇺 кальций, 🇪🇸 calcio

['w] 🇬🇧 calcium, كالسيوم, 🇫🇷 calcium, 🇩🇪 Kalzium, Calcium, 🇬🇷 ασβέστιο, 🇮🇱 קצ'ום, סידן, 🇮🇳 चूर्णातु, 🇮🇹 **calcio**, 🇯🇵 カルシウム, 🇷🇺 кальций, 🇪🇸 calcio

🇬🇧 calcium, Calcium in biology, كالسيوم, 🇨🇳 钙, 鈣営養, 🇫🇷 calcium, 🇩🇪 Calcium, 🇬🇷 Ασβέστιο, 🇮🇱 סידן, 🇮🇳 कैल्शियम, 🇮🇹 **calcio**, 🇯🇵 カルシウム, 🇷🇺 кальций, Кальций в живых организмах, 🇪🇸 calcio

Calcium regulation

Parathyroid glands

Babelfying the Multilingual Web: state-of-the-art Disambiguation and Entity Linking
Andrea Moro, Alessandro Raganato and **Roberto Navigli**

Calcio / Soccer in BabelNet 2.5

Meaning: soccer[1] • **ID:** bn:00006547n • **Type:** Concept [details] [explore]

Senses:

🇬🇧 soccer[1], association football[1]

🇫🇷 soccer, football, 🇬🇷 ποδόσφαιρο, 🇮🇱 כדורגל, 🇮🇹 **calcio**, pallone, football, 🇯🇵 アソシエーションフットボール, サッカー, フットボール, フートボール, 蹴球, 🇪🇸 fútbol

🇬🇧 Association football, 🇵🇸 لعبة كرة القدم, 🇨🇳 足球, 🇫🇷 Football, 🇩🇪 Fußball, 🇬🇷 Ποδόσφαιρο, 🇮🇱 כדורגל, 🇮🇳 फुटबॉल, 🇮🇹 **Calcio (sport)**, 🇯🇵 サッカー, 🇷🇺 Футбол, 🇪🇸 Fútbol

🇬🇧 Associated football, Outdoor Soccer, Soccer history, Voetbal, Association soccer, Association footballer, Footbal (soccer), Association football (soccer), Soccer Scoring, Soccer team, Recreational soccer, Fútbol, Sawker, 蹴球, Football (real), Fitba, Soccer football, Asocciation football, Football association, Football (association), Assocation football, Soccer (football), Association Football Club, Socker, Association Football, Ass. football, International club football, Football(soccer), Football (Soccer), World football, Soccer (Football), Association Football (soccer), Senior football, Football (Association), Futébol, Football (original), Soccor, Soccer, Assocciation football, Football (Association rules), The Method of Scoring, Balompie, サッカー, Balompié, Football (soccer), Futebol, Metric football, Association footballers, Rooball, اتحاد كرة القدم, 🇨🇳 Football, 英式足球, 踢足球, 足球运动, 🇫🇷 Équipe de football, Équipe de foot, Club de foot, Joueur de football, Football européen, Soccer, Le football, Footballeur, Contrôle (football), Tête (football), Fute balle, Association football, Football au Portugal, Contrôle de poitrine, Football Club, Football masculin, Le foot, Foot-ball, Footballeurs, Balle au pied, Footeux, Football association, 🇩🇪 Fußballspieler, Fussball, Fußballspielen, Fussballspieler, Fuathboi, Seitenaus, Fußballer, Männerfußball, Fussballer, Herrenfußball, 🇬🇷 Ποδοσφαιριστής,

Word Sense Disambiguation in a Nutshell

striker
(target word)

"Thomas and Mario are **strikers** playing in Munich"
(context)

knowledge

WSD
system

sense of target word
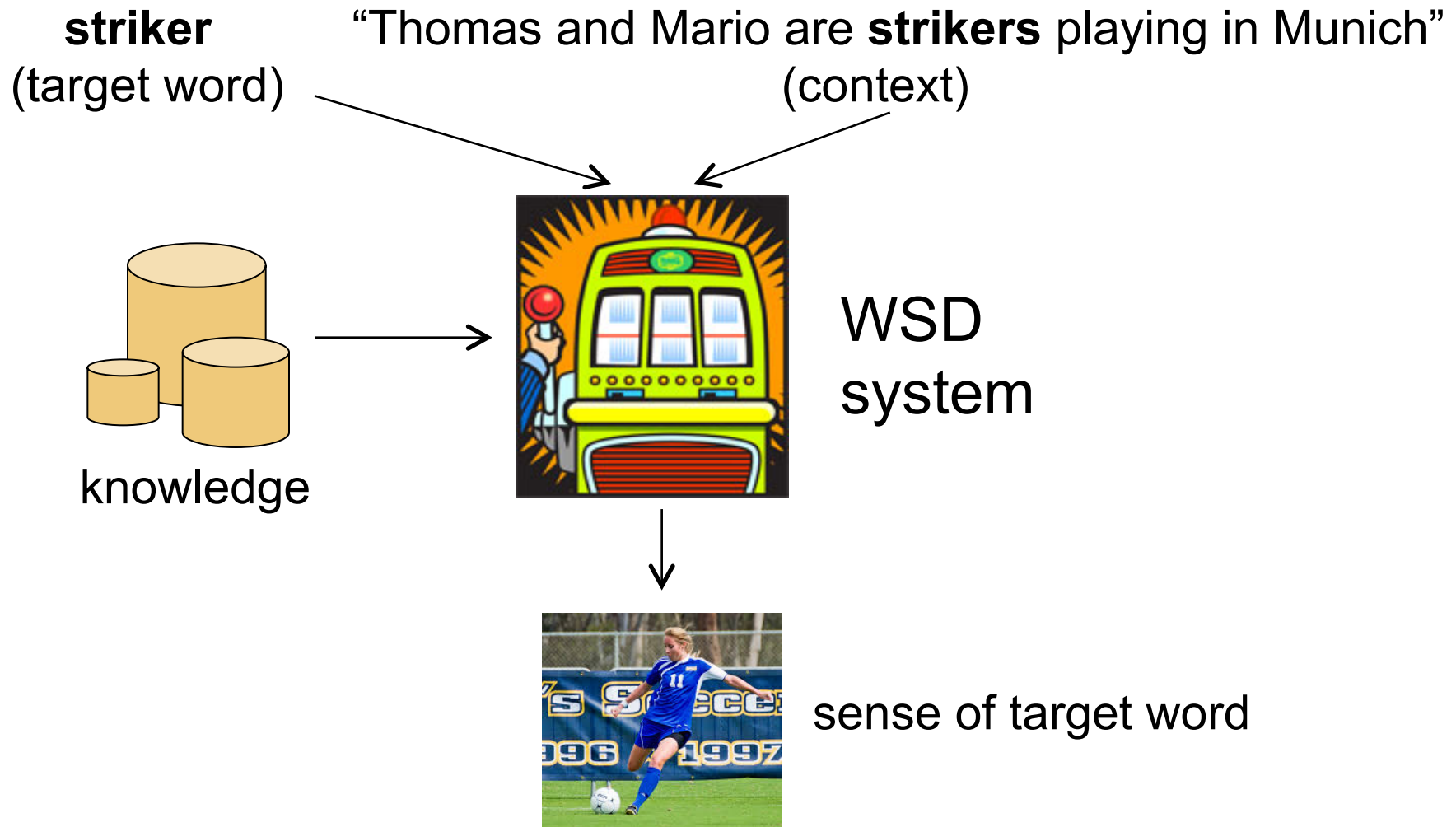
Entity Linking in a Nutshell

Thomas
(target mention)

"**Thomas** and Mario are strikers playing in Munich"
(context)

knowledge

Entity Linking system
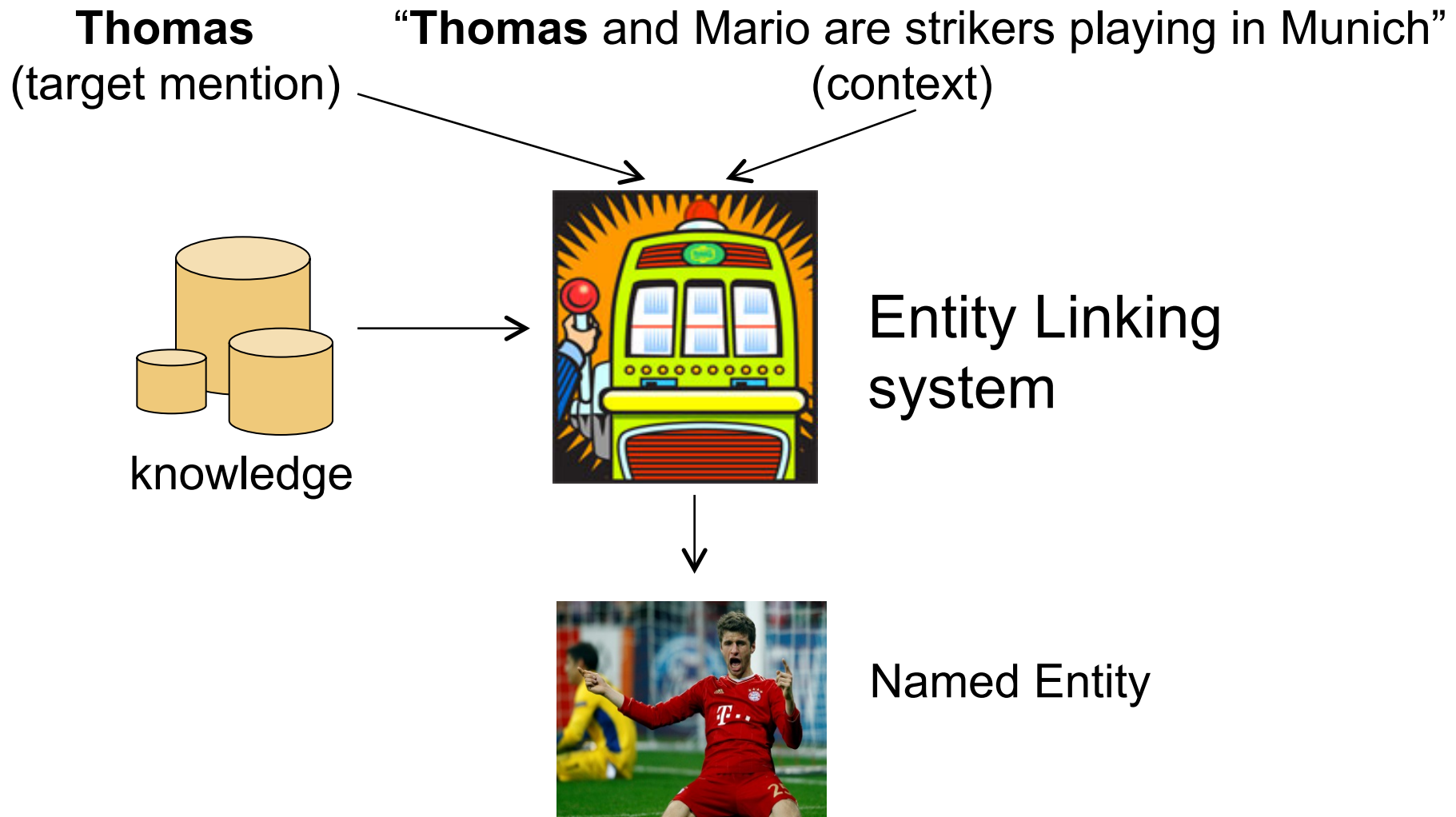
Named Entity

Disambiguation and Entity Linking together!

BabelNet is a huge **multilingual inventory**

for both word senses and named entities!

fy **B**abelfy .org

"Word sense disambiguation and entity linking together!"

Powered by BabelNet

Text to babelfy:

Enable partial matches: ☐ Language: English ▾ Babelfy!

Babelfy is an output of the MultiJEDI ERC Starting Grant No. 259234. Concept by Andrea Moro and Roberto Navigli. **Babelfy** and its **API** are licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License. (cc) BY NC SA For any commercial use, please contact us.

erc

Babelfying the Multilingual Web: state-of-the-art Disambiguation and Entity Linking
Andrea Moro, Alessandro Raganato and **Roberto Navigli**

Babelfy: A Joint approach to WSD and EL

Personalized PageRank is the state-of-the-art method for graph-based word sense disambiguation, however it **cannot** be run for each new input on **huge graphs**.

Idea: Precompute semantic signatures for the nodes!

Semantic signatures are the most **relevant nodes** for a given node in the graph computed by using **random walk with restart**

Andrea Moro and Alessandro Raganato and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. TACL

http://babelfy.org

Semantic Signatures: RWR

1. Start from one target vertex of the semantic netowork;

2. Randomly select a neighbor of the current vertex or restart at the target vertex;

3. Keep counting the hitting frequencies;

4. Take the most visited vertices.

Step 1: Calculate Semantic Signatures

striker

offside

athlete

sport

soccer player
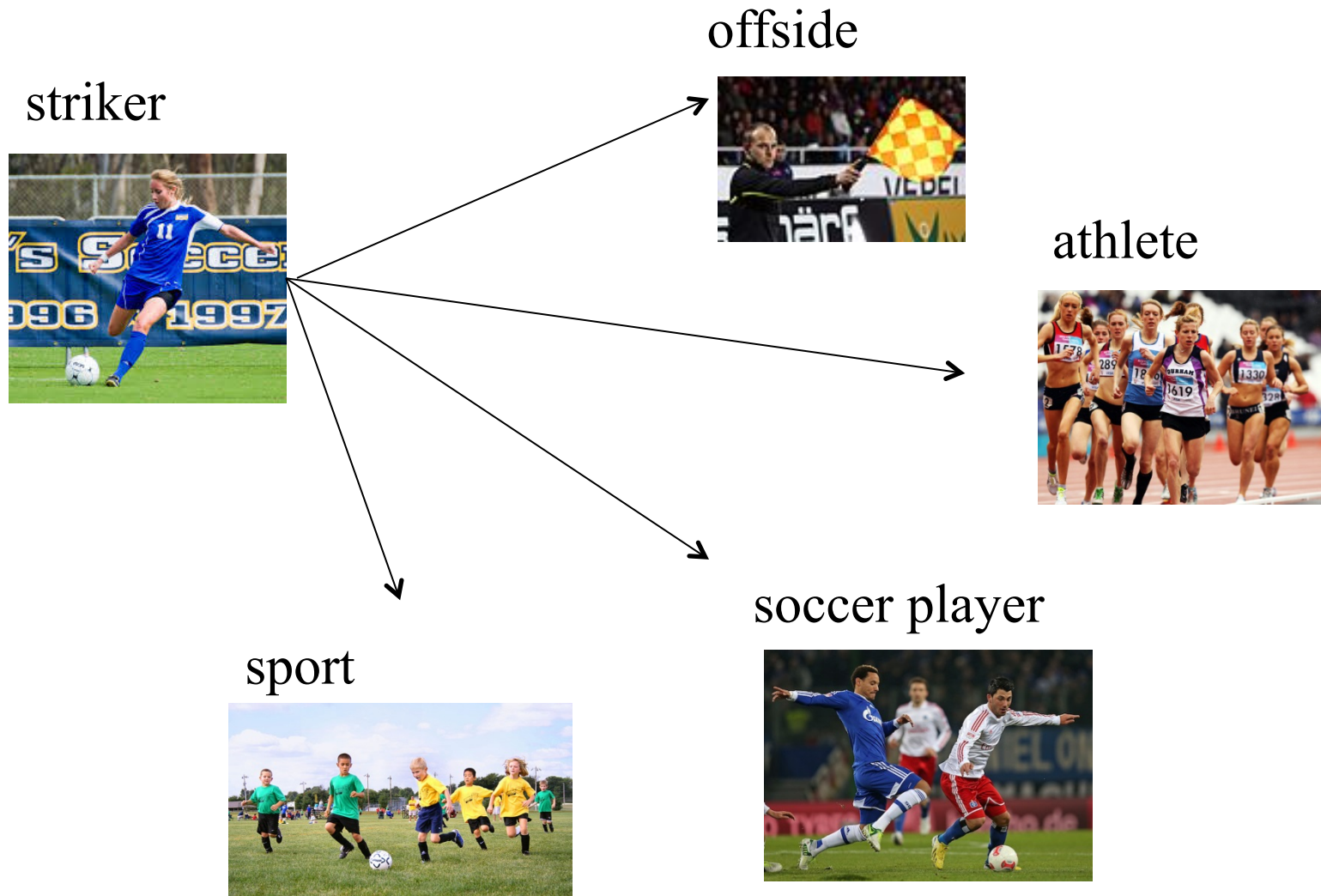
A Joint approach to WSD and EL

1. Given an input text select all the possible candidate meanings from BabelNet by matching mentions with BabelNet lexicalizations;

2. Connect all the candidate meanings by using semantic signatures;

3. Extract a dense subgraph containing semantically coherent candidates;

4. Select the most connected candidate for each fragment of text.

Step 2: Find all possible meanings of words in context

"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

Thomas (novel)

Mario Gómez

Striker (Movie)

Munich (Song)

Step 2: Find all possible meanings of words in context

"Thomas and Mario are strikers playing in Munich"

Seth Thomas

Mario (Character)

striker (Sport)

Munich (City)

Mario (Album)

Striker (Video Game)

Thomas Müller

Ambiguity!

FC Bayern Munich

Mario Gómez

Striker (Movie)

Thomas (novel)

Munich (Song)

Step 3: Connect all the candidate meanings

Thomas and **Mario** are **strikers** playing in **Munich**
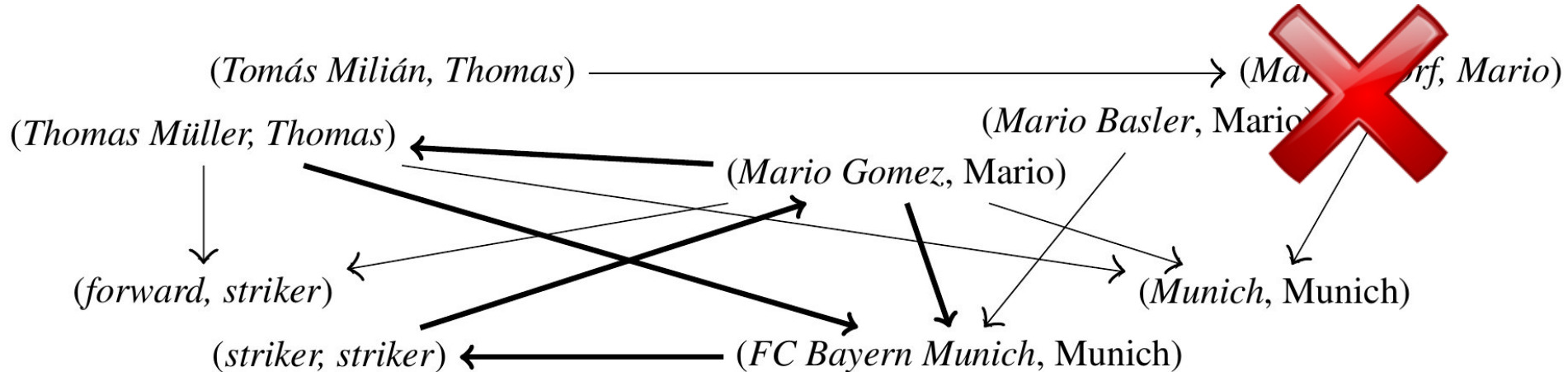
Step 4: Extract a dense subgraph

Thomas and **Mario** are **strikers** playing in **Munich**
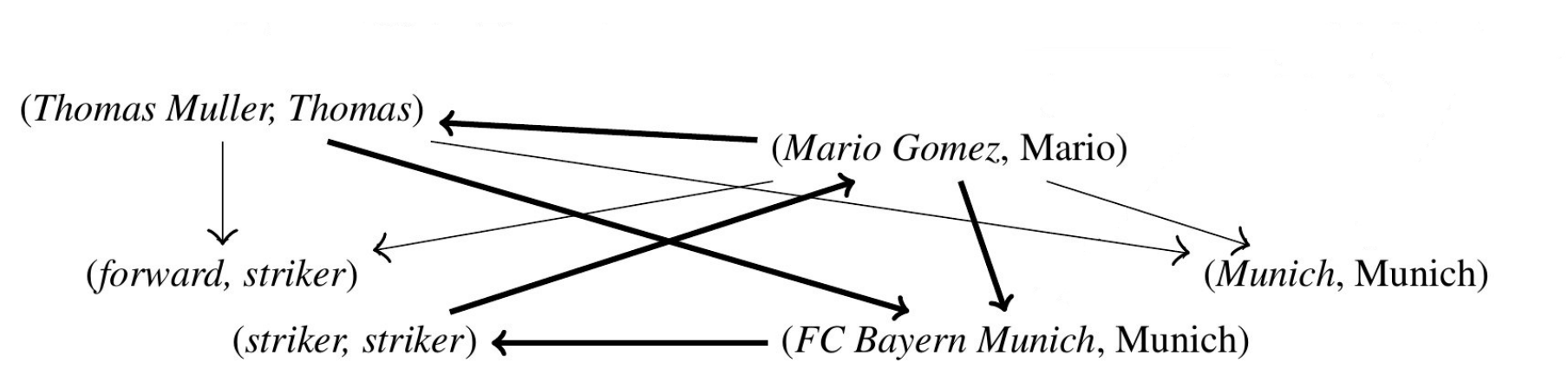
Step 5: Select the most reliable meanings

"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

Thomas (novel)

Mario Gómez

Striker (Movie)

Munich (Song)

Experimental Results: State-of-the-art multilingual disambiguation

	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

Table 1: F1 scores (percentages) of the participating systems of SemEval-2013 task 12 together with MFS, UKB w2w, IMS, our system and its ablated versions on the Senseval-3, SemEval-2007 task 17 and SemEval-2013 datasets. The first system which has a statistically significant difference from the top system is marked with ⋆ (χ^2, $p < 0.05$).

Experimental Results

System	F1
(Ponzetto and Navigli, 2010)	**85.5**
Babelfy	84.6
UoR-SSI	84.1
UKB w2w	83.6
NUS-PT	$\star$82.3
MFS	77.4
Babelfy unif. weights	85.7
Babelfy w/o dens. sub.	84.9
Babelfy only concepts	85.3
Babelfy on sentences	82.3

Table 2: F1 score (percentages) on the SemEval-2007 task 7. The first system which has a statistically significant difference from the top system is marked with $\star$ (χ^2, $p < 0.05$).

Experimental Results: State-of-the-art Entity Linking

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

Table 3: Accuracy (percentages) of state-of-the-art EL systems and our system on KORE50 and AIDA-CoNLL. The first system with a statistically significant difference from the top system is marked with ⋆ (χ^2, $p < 0.05$).

Stars boycott Beverly Hills Hotel over Brunei sharia move

Demonstrators protest over Brunei's new harsh Islamic penal code at the Beverly Hills Hotel

Stars of Hollywood are boycotting one of the area's most famous hotels because of a harsh Islamic penal code introduced by Brunei's government.

Comedians Jay Leno and Ellen DeGeneres are among the celebrities supporting the boycott against the Beverly Hills Hotel and other hotels

Related Stories

Brunei introduces tough Islamic law

Brunei profile

fy**B**abelfy

"Word sense disambiguation and entity linking together!"

Powe...

hotels
a building where travelers can pay for lodging and meals and other services

Stars
a widely known person; "he was a baseball celebrity"

sharia
the code of law derived from the Koran and from the teachings and example of Mohammed; "sharia is only applicable to Muslims"; "under Islamic law there is no separation of church and state"

erly Hills Hotel over Brunei ... famous hotels because of a...

o and Ellen DeGeneres are am...

boycott
refuse to sponsor; refuse to d... business with
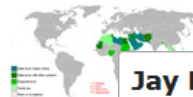
Jay Leno
James Douglas Muir "Jay" Leno is an American stand-up comedian and television host.

...rities
...ely known person; "he ... baseball celebrity"

Ellen DeGeneres
Ellen Lee DeGeneres is an American stand-up comedian, television host and actress.

Beverly Hills Hotel over Brunei shar...

one of the area 's...

introduced by Bru...

...res are among the ...

...e Beverly Hills Hotel an...

the Beverly Hills Hotel
The Beverly Hills Hotel, also called The Beverly Hills Hotel and Bungalows, is a hotel on Sunset Boulevard in Beverly Hills, California.

BBC

NEWS

6 May 2014 Last updated at 11:08 GMT

| News | Sport | Weather | Capital | Future | Shop | TV | Radio | More... |

Home | UK | Africa | Asia | Europe | Latin America | Mid-East | US & Canada | Business | Health | Sci/Environment | Tech | Entertainment | Video

▶ Thai crisis explained in 60 seconds

Profile: Yingluck Shinawatra

Thailand to hold fresh election

Q&A: Thailand protests

Hope for boost on Ukraine elections

European foreign ministers seek to build support for this month's presidential election in Ukraine, where troops are battling pro-Russian separatists.

LIVE Russia's Lavrov holds news briefing

▶ Ukraine soldiers killed in Sloviansk

Odessa images

How did Odessa's fire happen?

Coca-Cola to remove drinks ingredient

Coca-Cola is to remove an ingredient which has been linked to negative health effects from some of its drinks brands by the end of this year.

Can big food firms get us to eat more healthily?

Campaigners vow to cut sugar in food

Features

Migrants

Trapped in th
workers, drea

Toppling a

Does China's
down the for

Living ou

Young Europ
hand-to-mou

Too old?

The 93-year-
India's electi

Haunted

Why South A
veterans say

31-0

The story be
football team

Muscular

![Babelfy logo — "Word sense disambiguation and entity linking together!"]

Powered by BabelNet

Text to babelfy:

European foreign ministers seek to build support for this month's presidential election in Ukraine, where troops are battling pro-Russian separatists.

Enable partial matches: ☐ Language: English ▼ Babelfy!

European foreign ministers seek to build support for this month 's

foreign ministers
a government minister for foreign relations

kraine , where troops are battling pro-Russian

separatists .

te-of-the-art Disambiguation and Entity Linking
and **Roberto Navigli**

"Word sense disambiguation and entity linking together!"

fyBabelfy

Powered by BabelNet

Text to babelfy:

The world's largest beverage-maker, Coca-Cola, plans to rem[move] some of its drinks brands by the end of this year, followin[g]

Brominated vegetable oil, or BVO, is found in Coca-Cola fru[it] and Powerade.

An element of the additive, also found in flame retardants,

drinks
any liquid suitable for drinking; "may I take your beverage order?"

Enable partia[l]

The world 's larges[t]

Fanta
Fanta is a global brand of fruit-flavored carbonated soft drinks from the Coca-Cola Company.

flame retardants
Flame retardants are chemicals used in thermoplastics, thermosets, textiles and coatings that inhibit or resist the spread of fire.

[en]ds by the end   of this

[vege]table oil , or BVO , is

Fanta and Powerade .

[entity l]inking

Next Goal Wins for 'world's worst football team'

By Kev Geoghegan
Arts and entertainment reporter, BBC News

Goalkeeper Nicky Salapu is the only remaining player from the 31-0 hammering by Australia

In 2001, Australia set a world record for the largest ever victory in an international football match, netting 31 goals without a single breach of their own defence.
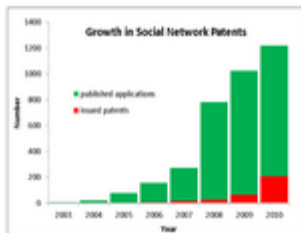
Related Stories

Babelfy

"Word sense disambiguation and entity linking together!"

Powered by BabelNet

social media
Social media employ web - and mobile-based technologies to support interactive dialogue and "introduce substantial and pervasive changes to communication between organizations, communities, and individuals."

Growth in Social Network Patents

Twitter
Twitter is an online social networking service and microblogging service that enables its users to send and read text-based messages of up to 140 characters, known as "tweets".

o con un cinguettio senza
mazonBasket nel Regno
isce gli account Twitter
del suo artista favorito,
u Amazon.com, basterà
rt o #AmazonBasket per

Babelfy!

Amazon.com
Amazon.com, Inc. is an American multinational electronic commerce company with headquarters in Seattle, Washington, United States.

amazon.com

Amazon   con un   cinguettio   senza   lasciare   il

social med   no Unito   , i   due   Paesi   dove   finora   il

servizio   e   disponibile   ccount   Twitter   e Amazon degli   utenti   . Come   funziona   ? Quando un

cliente   legge   un   tweet   o   , di una   marca   o di un   amico   con un   link   con un

prodotto   su   Amazon.co   sta   un tweet   seguito   dall '   hashtag   #AmazonCart o

                                              rio   carrello   su   Amazon .

Thanks or…

merci

(grazie)

mylider

Roberto Navigli

Linguistic Computing Laboratory

http://babelnet.org
http://babelfy.org