Multilingual text mining and Machine Translation activities carried out at the EC's Joint Research Centre (JRC)

Marco Turchi
& the JRC's *OPTIMA* team – Open Source Text Information Mining and Analysis

Technical details and publications: http://langtech.jrc.ec.europa.eu/
Applications: http://emm.newsbrief.eu/overview.html

JRC
EUROPEAN COMMISSION

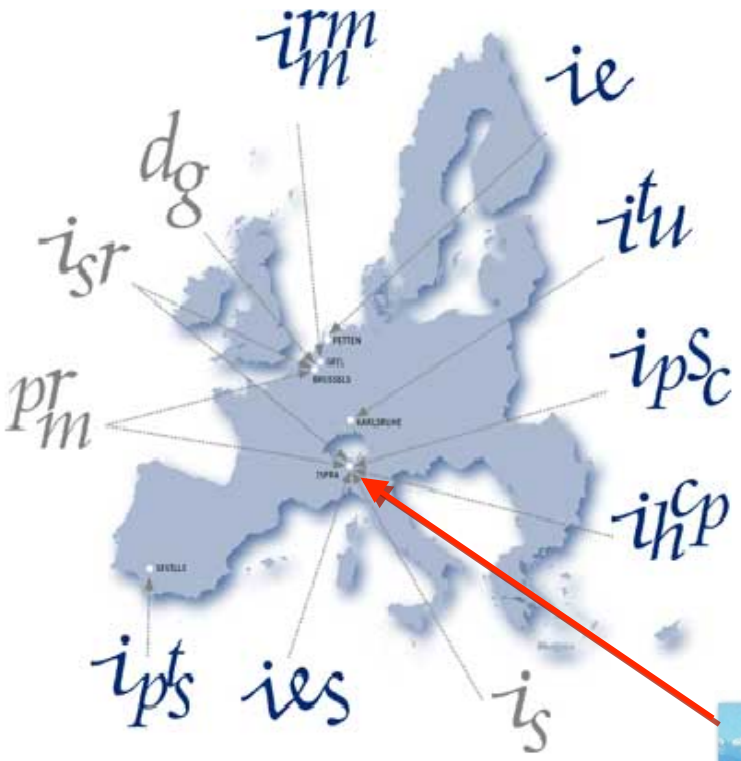
Outline

- Joint Research Center - Who we are
- Multilinguality:
  - Language-independent algorithm
    - Case Study:

      Multilingual named entity recognition and variant mapping
  - Machine Translation
    - Case Study

      Optima Machine Translation Service

Joint Research Centre - Who we are



BRUSSELS (BE)
The Directorate General (**DG**)
The Institutional and Scientific Relations Directorate (**ISR**)
The Programme and Resource Management Directorate (**PRM**)

GEEL (BE)
The Institute for Reference Materials and Measurements (**IRMM**)

KARLSRUHE (DE)
The Institute for Transuranium Elements (**ITU**)

ISPRA (IT) Download the Ispra site Brochure (English - Italian)
The Institute for the Protection and Security of the Citizen (**IPSC**)
The Institute for Environment and Sustainability (**IES**)
The Institute for Health and Consumer Protection (**IHCP**)
The Ispra site Directorate (**IS**)

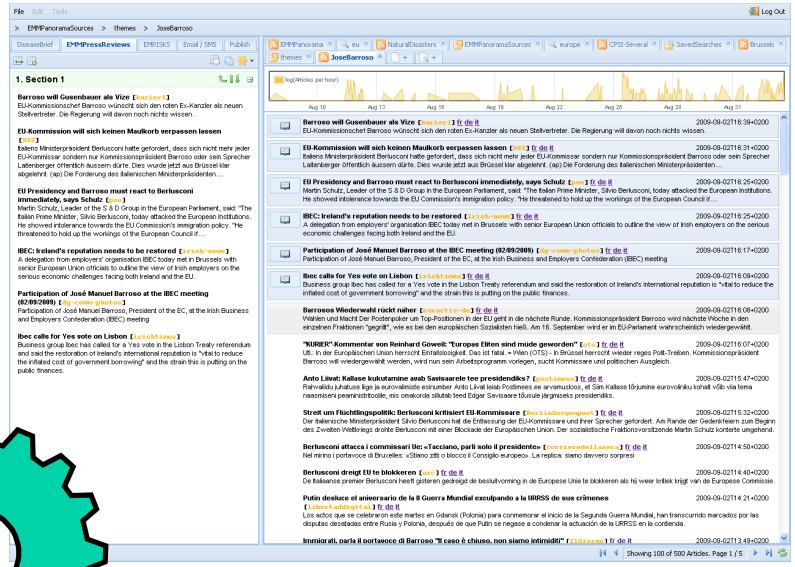
PETTEN (NL)
The Institute for Energy (**IE**)

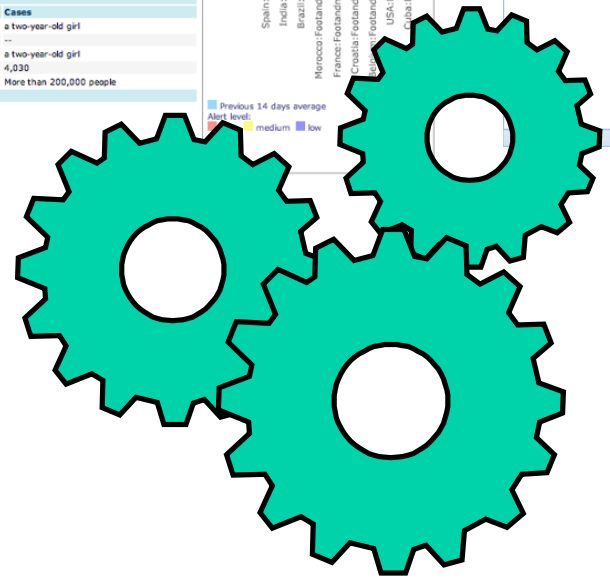
SEVILLE (E)
The Institute for Prospective Technological Studies (**IPTS**)

"The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national."

Europe Media Monitor - EMM

NewsBrief, Medisys, BlogBrief





NewsDesk Service (a.k.a. RNS)
Editorial Interface

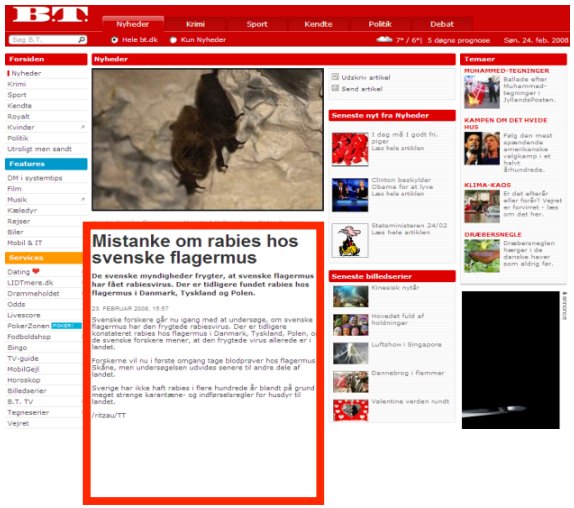
EMM Open Source Information Monitoring Engine

Europe Media Monitor (EMM)

- **~ 3,400 Sources** (world-wide, with focus on Europe)
  - news sources (web portals)
  - specialist medical sites
  - commercial newswires
  - 24/7, updated every 10 minutes

- **~ 100,000 articles / day in ~ 50 languages**

- Converts dirty html with adverts, menus, html tags, 'related stories', etc. into clean and standardised UTF-8 encoded RSS format.



- Articles are fed into the various EMM applications/algorithms.

Europe Media Monitor (EMM)

- **~100,000 articles per day… in ~ 50 languages.**
- **Fast!**

- Research Area:
  - Multilingual Person Name Recognition
  - Name Variation Matching
  - News Clustering
  - Cross-lingual Linking of News Clusters
  - Document Summarization
  - Sentiment Analysis
  - Automatic Event Extraction
  - Co-reference Resolution
  - Document Classification
  - Machine Translation
  - …

The EMM Family of News Analysis Applications

- **NewsBrief**: current state of affairs, *breaking news* detection in real time
- **MedISys**: focusing on *health-related* news
- **NewsExplorer**: long-term, *cross-lingual* news analysis and *people and organization* monitor
- **EMM-Labs**: various data *visualization* and advanced *text processing* tools

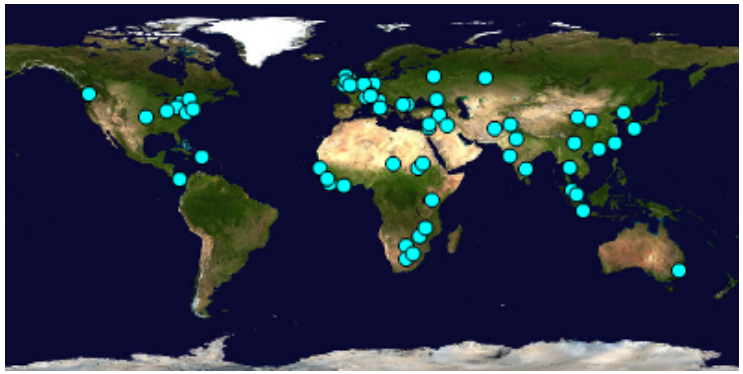
http://emm.newsbrief.eu/

JRC
EUROPEAN COMMISSION

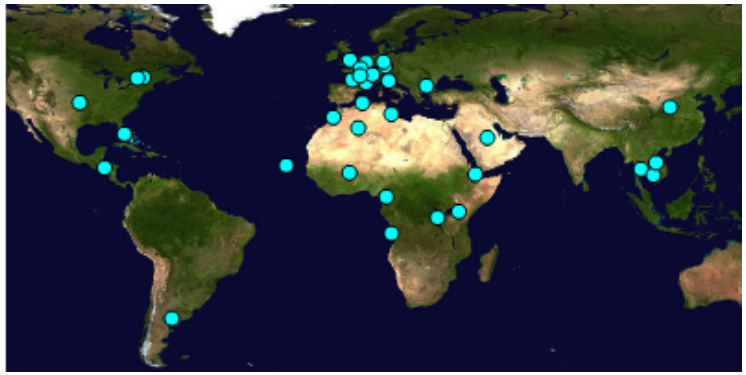
Why Multilinguality?

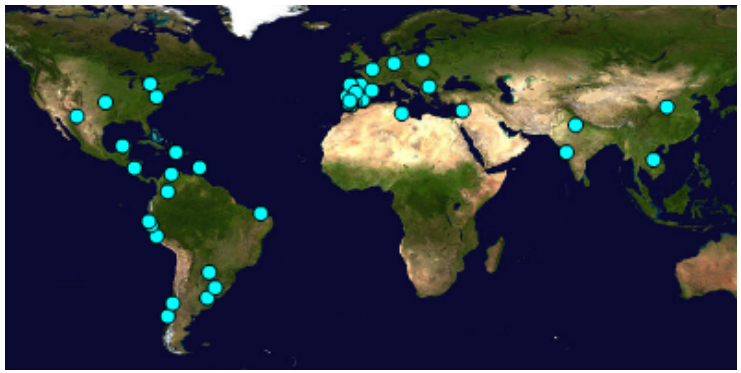
Locations mentioned in medical articles across languages

Italian - German

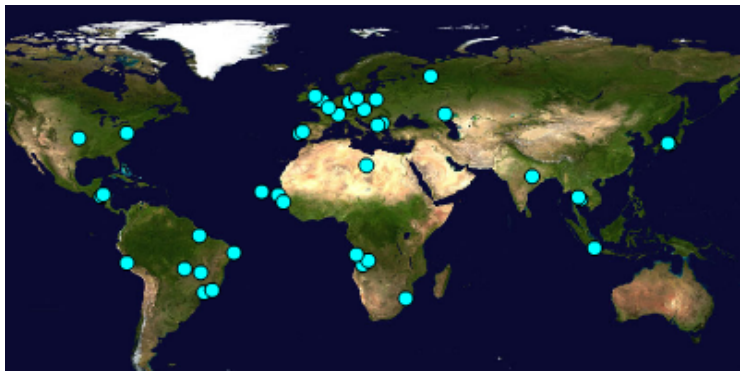


English - French





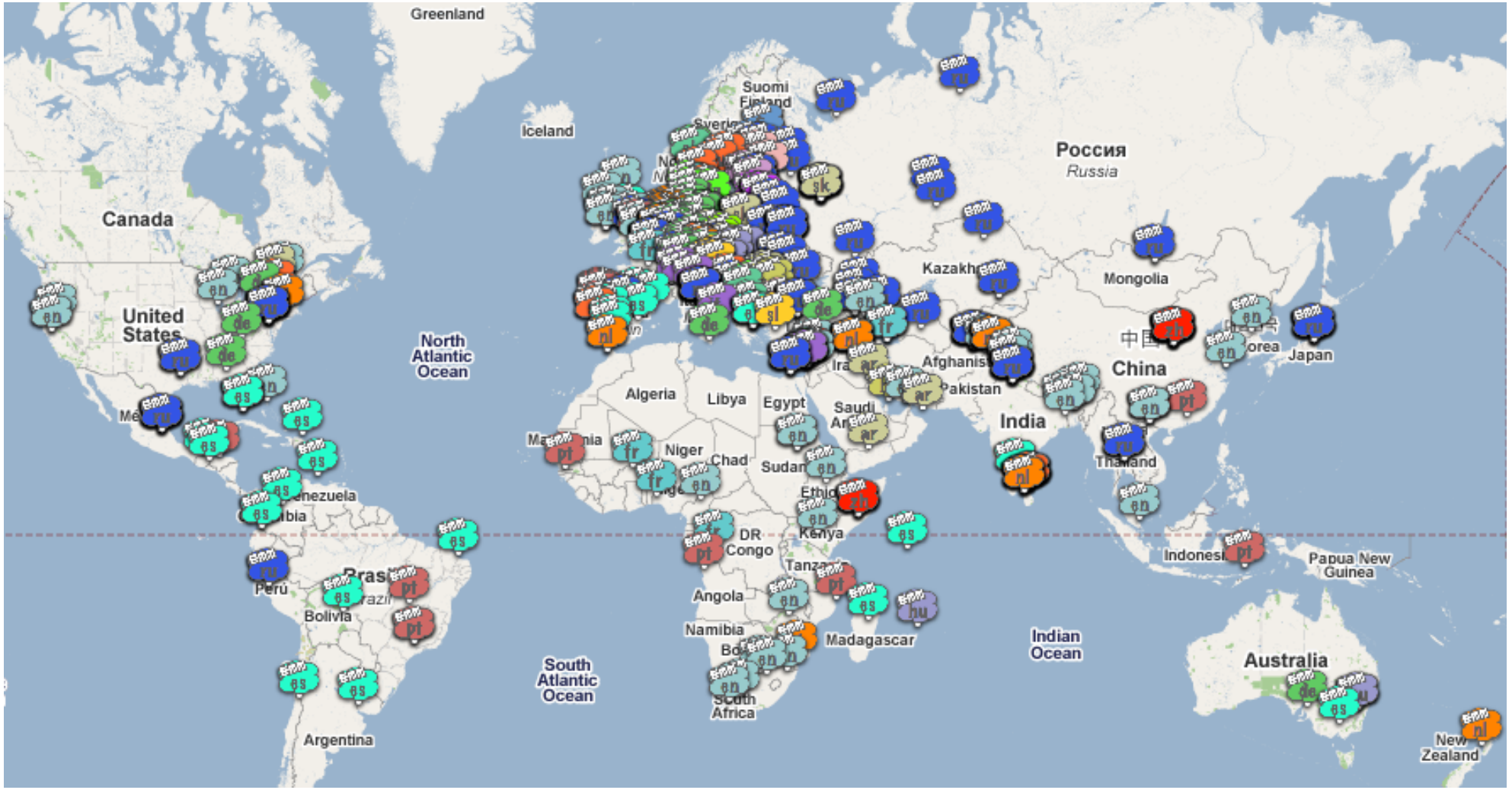
Spanish - Portuguese



JRC
EUROPEAN COMMISSION

Live Cluster Map

Display of latest geo-located news clusters



How to address multilinguality?

- **Language-independent** algorithms (as much as possible).

- **Machine Translation** :
  - **Direct**:
    - Translate documents to a common language.
  - **Indirect**:
    - Use the translation engine to improve language independent algorithms.

JRC
EUROPEAN COMMISSION

Outline

- Joint Research Center - Who we are
- Multilinguality:
  - Language-independent algorithm
    - Case Study:

      Multilingual named entity recognition and variant mapping
  - Machine Translation
    - Case Study

      Optima Machine Translation Service

Multilingual name recognition and variant mapping

en death of former Prime Minister ***Rafik Hariri***, blamed by many opposition

es asesinato del exprimer ministro ***Rafic al-Hariri***, que la oposición atribuyó

fr l'assassinat de l'ex-dirigeant ***Rafic Hariri*** et le départ du chef de la diplom

nl na de moord op oud-premier ***Rafiq al-Hariri*** gingen gisteren bijna een

de libanesischen Regierungschef ***Rafik Hariri*** vor einem Monat wichtige B

sl danjega libanonskega premiera ***Rafika Haririja***. Libanonska opozicija si

et möödumisele ekspeaminister ***Rafik al-Hariri*** surma põhjustanud pommipl

ar اغتيال رئيس الوزراء السابق ***رفيق الحريري*** بأياد يهودية وما حدث سابقا

ru Бывший премьер-министр ***Ливана Рафик*** Харири, который

NER: Known person + organisation names

- **Lookup of known names** from database
  - Currently about 1,150,000 names
  - + 225.000 variants
  - Pre-generate morphological variants (Slovene example):

Tony(a|o|u|om|em|m|ju|jem|ja)?\s+**Blair**(a|o|u|om|em|m|ju|jem|ja)

Tony Blair / Tonyju Blairu / Tony Blairt / Tonijs Blērs / Тони Блэра / توني بلير / Tony Blairi / طوني بلير

NER: Known person + organisation names

- **Guessing names** using empirically-derived lexical patterns
  - **Trigger word(s)** + ***Uppercase Words***
    (+ name particles: von, van, de la, abu, bin, …)
    - **President, Minister, Head of State, Sir, American**
    - **"death of", "[0-9]+-year-old", …**
    - Combinations: "**56-year-old former prime minister *Kurmanbek Bakiyev***"
  - First name lists (**John, Jean, Hans, Giovanni, Johan,** …)

- Identification of a current average of 608 unknown names per day
- 83 of those are automatically merged with known names

- Use **bootstrapping** to produce a trigger word list for a new language
  - Start with small initial trigger word list or lists of known names
  - Produce frequency list of contexts of known names
  - Manual selection

NER: Inflection of trigger words

- Inflection of trigger words for person names, using regular expressions (Slovene example):
  - **kandidat**(a|u|om)?
  - **legend**(a|e|i|o)
  - **milijarder**(ja|ju|jem)?
  - **predsednik**(a|u|om|em)?
  - **predsednic**(a|e|i|o)
  - **ministric**(a|e|i|o)
  - **sekretar**(ja|ju|jom|jem)?
  - **diktator**(ja|ju|jem)?
  - **playboy**(a|u|om|em)?

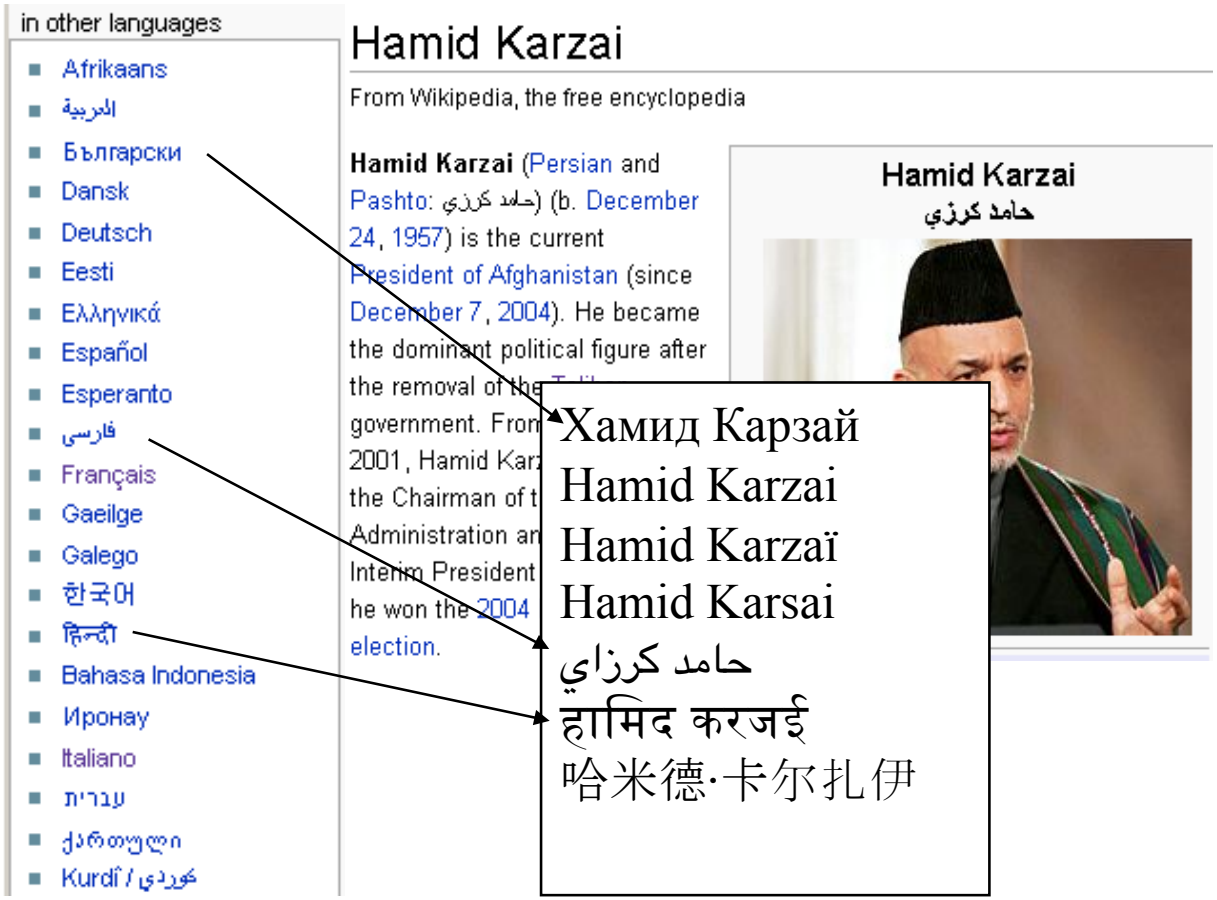
+ ***uppercase words***

… verskega **voditelja** ***Moktade al Sadra*** je z notranjim …

= ***Muqtada al-Sadr*** (ID=236 in the DB)

Name Variants

- Adding names (and images) from Wikipedia
- Merging name variants
  - **Transliteration**
  - **Normalisation**
  - **Similarity measure**



Name transliteration

- Currently, EMM NewsExplorer transliterates from Arabic, Farsi, Greek, Russian and Bulgarian

- **Transliterate each character**, or sequence of characters, by a Latin correspondent
  - ψ => ps
  - λ => l
  - μπ => b

  - Examples of transliterations:
    - Κόφι Ανάν, Greek → ***Kofi Anan***
    - Кофи Аннан, Russian → ***Kofi Annan***
    - Кофи Анан, Bulgarian → ***Kofi Anan***
    - कोफी अन्नान, Hindi → ***Kofi Anan***

Name normalisation – Why? (1)

- **Transliteration rules** depend on the target language, e.g.
  Владимир Устинов (Russian)
  - ***V***ladimir ***U***stino***v*** (English)
  - ***W***ladimir Ustino***w*** (German)
  - Vladimir ***Ou***stinov (French)

- Various ways to represent the **same sound**: sh, sch, ch, š
  e.g.
  - Ba***š***ar al Assad
  - Ba***sch***ar al Assad
  - Ba***ch***ar al Assad

Names
Bashar al-Assad (Eu,yo)
بشار الأسد (ar)
Bashar Assad (Eu,sv)
Bachar al-Assad (es,pt)
Bachar el-Assad (fr)
Bashar al Assad (da,pt)
Baschar al-Assad (de,nl)
Башар Асад (bg,ru)
Beşar Esad (tr)
Beşşar Esad (tr)
Bashar Al-Assad (en,ro)
Bachar Al-Assad (es,pt)
Bashar Al Assad (en,tr)
Bachar al Assad (es,pt)
Baschar el Assad (de)
Bachar al Asad (es,pt)
Baschar al Assad (de,es)
Başar Esad (tr)
Bašar al Asad (sl)
Bashar el Assad (de,it)
Bashar al Asad (es)

Name normalisation – Why? (2)

- **Diacritics** are often omitted, e.g.
  - Wałęsa → ***Walesa***
  - Saïd → ***Said***
  - Schröder → ***Schroder***
  - Skarsgård → ***Skarsgard***
  - Jørgen → ***Jorgen***

→ **Edit distance** is large for naturally occurring word variants.

Name normalisation (3) - ~ 30 Rules

- **Latin normalisation:**
  - accented character → non-accented equivalent
  - double consonant → single consonant
  - ou → u
  - “ al-” →
  - wl (beginning of name)→ vl
  - ow (end of name) → ov
  - ck → k
  - ph → f
  - ž → j
  - š → sh
  - x → ks
- **Remove vowels**

Malik al-Saïdoullaïev
Malik al-Saidoullaiev
Malik al-Saidoulaiev
Malik al-Saidulaiev
Malik Saidulaiev
… ***mlk sdlv***

Name	Normalised form
Mohammed Siad Barre, Mohamed Siad Barré, Мохаммед Сиад Барре, محمد سياد باري	**mhmd sd br** (mohamed siad bare)
Mahmoud Ahmadinejad, Mahmūd Ahmadīnežād	**mhmd hmdnjd** (mahmud ahmadinejad)
Сергей Куприянов, Sergei Kupriyanov, Sergei Kuprianow, Sergueï Kouprianov	**srg kprnv** (sergei kuprianov)
Ban Ki-moon, Ban Ki Moon, Пан Ги Мун	**bn k mn** (ban ki mun)

Similarity measure for name merging

- To compare ~600 new names every day with ~1,375,000 known name variants

→ **Only if** the transliterated, normalised form with vowels removed is identical

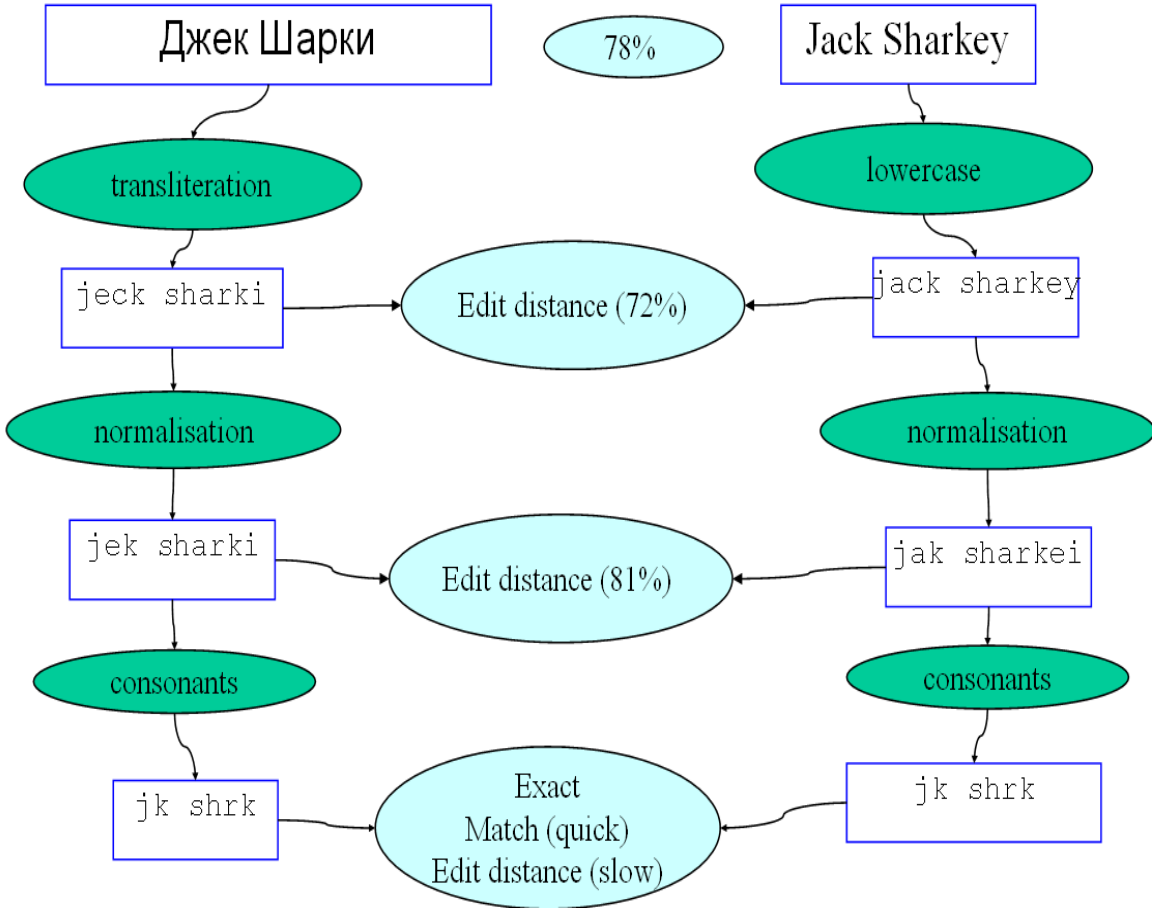
- Calculate edit distance variant similarity using two different representations:

20%

+

80%

Condition:



Name variants found in 16 hours of multilingual news analysis (25.3.2011)

2190. Muammar Gaddafi

2011-03-24T23:59+0200

Muammar Gaddafi / Muammar al-Gaddafi / معمر القذافي / Mouammar Kadhafi / Moammar Gadhafi / Муамара Каддафи / Муаммара Каддафи / Muamar Gadafi / Muammar Gheddafi / 卡扎菲 / Muamar al Gadafi / Muamar el Gadafi / Muammara Kadafiego / Muammer Kaddafi / Muammar al-Ghadhafi / Muammar Kadhafi / Муамар Кадафи / Muammara Kaddáfí / Moamer Kadhafi / Muammar al Gaddafi / Moamer Gaddafi /

Muammar al Gaddafi / Moamer Gaddafi / Muammar Gaddafin / Moammar Kadhafi / Muammar el Gaddafi / Muammar Qaddafi / Moamerja Gadafija / Муамару Каддафи / Muammar Kaddafi / Muamaro Gaddafi / Муаммару Каддафи / Muamar Kadafi / Муамар Каддафи / Moammar Qaddafi / Moamar Gadafi / Muammar Kaddáfí / Moamer Gadafi / Muammar Kadafi / لمعمر القذافي / Muammar Khadafi / Muammar Gadhafi / Muammar al Gadafi /

Муаммар Каддафи / Muammar Gadafy / Moammer Kadhafi / Muammar el-Qaddafi / Muamar Kadhafi / Moammar Kadafi / معمر قذافي / Муаммаром Каддафи / Moammar Gaddafi / Muammar Khaddafi / Muamaro Kadhafi / Muamar Khadafi / Muammar Khadaffi / Muammer Gaddafi / Muammaro Gaddafi / Moamar el Kadafi / Muhammar Gheddafi / Mouammar al-Kadhafi / Muammar Kadhaffi / Muamar El Gadafi / Moamar

Gaddafi / Muammar el-Gaddafi / Muammar al-Gadafi / Mouammar Kaddafi / Muammarem Kaddáfí / Muammar Gadafi / Muammar al-Gadhafi / Muammar al-Gaddafi / معمر القذافي / Muammar al-Qadhafi / Muamarui Gaddafi / Muammar Gadaffi / Muamaro Kadafio / Muamar Gaddafi / Muammar Qadhafi / Muammaro Gaddafio / Muamarui Gaddafiui / Muammar al-Gadaffi / Muamar al Gaddafi / Муамаром Каддафи / Muammar Ghadhafi /

Moamerju Gadafiju / Muammarowi Kadafiemu / Moammar Khadafy / Moammer Gaddafi / Muammar El Gadaffi / מועמר קדאפי / Mouamar Kadhafi / Mouammar El Kadhafi / Muammar Al-Gaddafi / Mummar Gaddafi / Muamar Gadaffi / Moammar Ghadafi

Person name recognition and variant merging – Result

Name variants

Ali Larijani (Eu,en)
Ali Laridschani (de)
Ali Lariyani (es)
Ali Laridžani (sl)
علي لاريجاني (ar)
Alí Larijani (es)
Ari Larijani (en)
Alí Lariyani (es)
Али Лариджани (ru)
Alì Larijani (it)
Ali Laridjani (fr)
Ali Larjani (sv)
Ali Lariani (it)
Ali Laryani (es)
Ali Laranjani (pt)
Ali Larigani (en)
Ali Larinjani (nl)
Al Larijani (nl)
Ali Ardashir Larijani (en)
Ali A. Larijani (en)
Ali Larejani (it)
Ali Larichani (es)
علی اردشیر لاریجانی (fa)
Ali Larijan (en)
Al Laridschani (de)
Ali Ardeshir Larijani (en)
Ali Laridziani (en)

Trigger words

negotiator (en - 824)
chefunterhändler (de - 470)
iraniano (it,pt - 311)
negociador iraniano (pt - 102)
iranien (fr - 129)
iranian negotiator (en - 87)
pogajalec (sl - 90)
unterhändler (de - 111)
iraní (es - 121)
iraanse onderhandelaar (nl -

A subset of the most important name variants is available:
http://langtech.jrc.ec.europe.eu/JRC-Names.html

JRC
EUROPEAN COMMISSION

Outline

- Joint Research Center - Who we are

- Multilinguality:
  - Language-independent algorithm
    - Case Study:

      Multilingual named entity recognition and variant mapping

  - Machine Translation
    - Case Study

      Optima Machine Translation Service

Machine Translation @ JRC

- **Machine Translation (MT) to directly or indirectly address multilinguality.**

- Two ways of using Statistical Machine Translation:
  - **Direct**:
    - Translate documents to a common language.
  - **Indirect**:
    - Use the translation engine to improve language-independent algorithms.

Optima Machine Translation System

- EMM gathers around **100k articles** from daily news in **50 languages**.

- Main requirements:
  - Help to **understand the content** of the news articles;
  - **Fast**;
  - **Language-independent**.

JRC
EUROPEAN COMMISSION

Key Questions

- Five key questions when you build a translation service for news and not only:
  1. **Which is the most suitable SMT system for our environment and requirements**?

JRC
EUROPEAN COMMISSION

Which is the most suitable SMT system?

- Main requirements are **translation speed** and **quality**
- Options:
  - PBSMT
  - Hierarchical
  - Syntax
- Train and tune the three systems using German-English Europarl V4 and test on news.

System	**Time (sec./sent.)**	**Quality (Bleu Score)**
PBSMT	**1.02**	18.09
Hierachical	4.5	**18.31**
Syntax	49	17.62

- Many source languages:
  - **No POS taggers, Syntactic Parsers** (maybe on the target side)
- System combination requires to **translate twice the same sentence**.

Key Questions

- Five key questions when you build a translation service for news and not only :
  1. Which is the most suitable SMT system for our environment?
     - ***PBSMT***
  2. **How to translate Named Entities in news**?

How to translate Named Entities in news?

- To understand the news content, named entities need to be correctly translated:
  - Source:
    - "*le ministre **bruno le maire** a presenté à **londres** les détails d' un plan gouvernemental de lutte contre les algues vertes .* "
  - Translated Sentence:
    - "*minister **bruno mayor** has presented in **london** the details of a government plan to fight against the green algae .* "
- Difficult to understand who has presented!!

How to translate Named Entities in news?

- **Isolates named entities** (person, location and organization) and **suggests their correct translation** in English to the SMT system:
  - Source:
    - "*le ministre* ***<bruno le maire:bruno le maire>*** *a presenté à* ***<londres:london>*** *les détails d' un plan gouvernemental de lutte contre les algues vertes .* "
  - Translated Sentence:
    - "*minister* ***bruno le maire*** *has presented in* ***london*** *the details of a government plan to fight against the green algae .* "
- **English translations** are obtained from our database.
- In case of more than one English translations, the most frequent is selected (problem during evaluation: EU or European Union?)

JRC
EUROPEAN COMMISSION

Key Questions

- Five key questions when you build a translation service for news and not only:
  1. Which is the most suitable SMT system for our environment?
     - ***PBSMT***
  2. How to translate Named Entities in news?
     - ***Suggests their correct translation to the SMT system***
  3. **How to deal with different language styles in news**?

How to deal with different language styles in news?

1. "*25 civilians dead as Taliban intensifies attacks in Afghanistan*"

2. "*Twenty-five people were killed in the latest round of Afghan violence this week in Kabul.*"

Comments?

How to deal with different language styles in news?

- Title and Content have a different language style
  - **Title** contains **more gerund** verbs, **no or few linking verbs**, **prepositions** and **adverbs**.
    - "*25 civilians dead as Taliban intensifies attacks in Afghanistan*"
  - **Content** contains **preposition**, **adverbs**, and **different verbal tenses**.
    - "*Twenty-five people were killed in the latest round of Afghan violence this week in Kabul.*"
- We trained two instances of the translation system:
  - in the "**title system**", parameters are optimized using parallel and monolingual titles.
  - in the "**content system**", parameters are optimized using normal parallel and monolingual sentences.

Optima Machine Translation System

- Some Results for De-En
  - Bleu score

	SMT_{Title}	$SMT_{Content}$
$Test\text{-}Set_{Title}$	***0.3706***	0.3511
$Test\text{-}Set_{Content}$	0.1768	***0.1945***

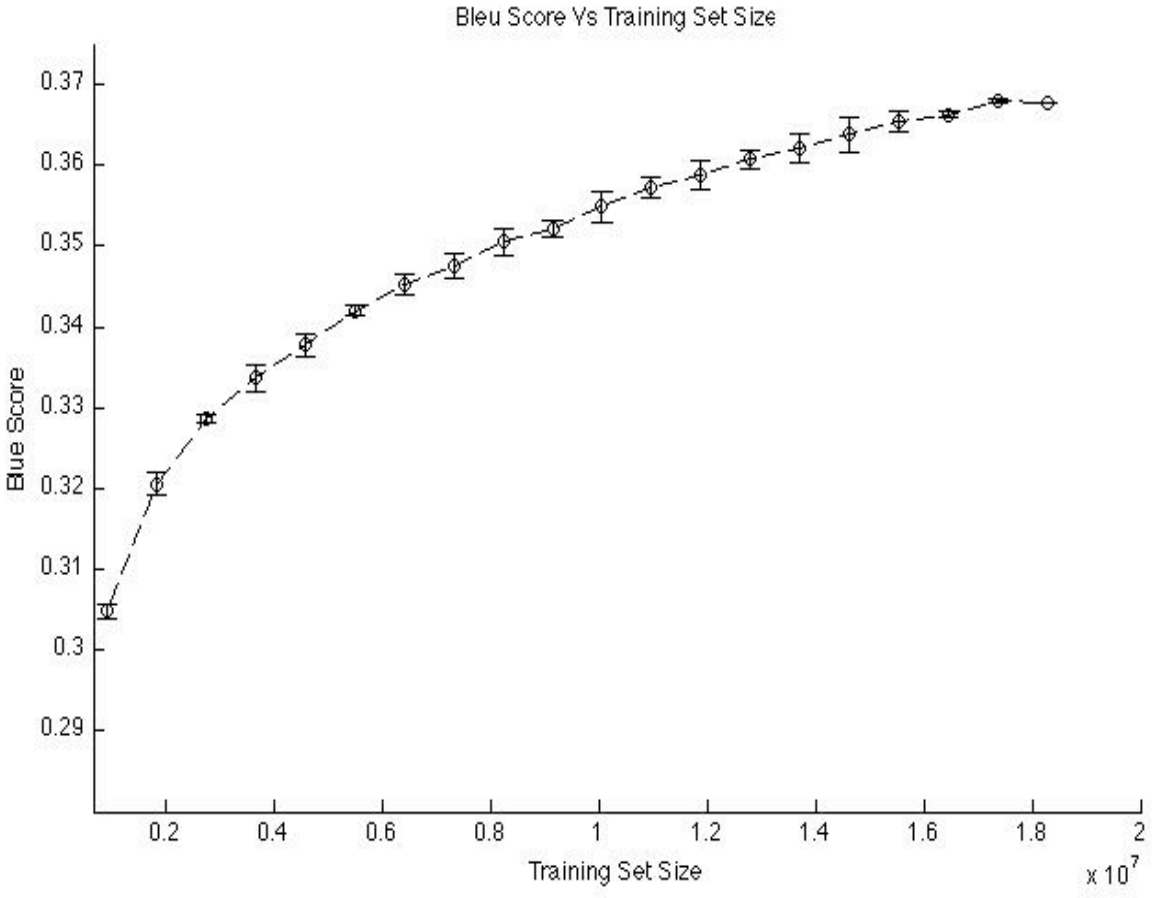
  - Similar results are obtained for the other language pairs.
  - Title data is gathered using Google Translate.

JRC
EUROPEAN COMMISSION

Key Questions

- Five key questions when you build a translation service for news and not only:
  1. Which is the most suitable SMT system for our environment?
     - ***PBSMT***
  2. How to translate Named Entities in news?
     - ***suggests their correct translation to the SMT system***
  3. How to deal with different language styles in news?
     - ***Title and Content customization***
  4. **Which training data can we use**?

Which training data can we use?



- Learning curve
- Use Giga corpus (> 18 M sentence pairs) for training and testing.

- Addition of massive amounts of data **from the same distribution** result in small improvements in the performance:
  - same results using different language pairs (**es-en, zh-en**) and SMT system (**Moses, Portage**).

JRC
EUROPEAN COMMISSION

Which training data can we use?

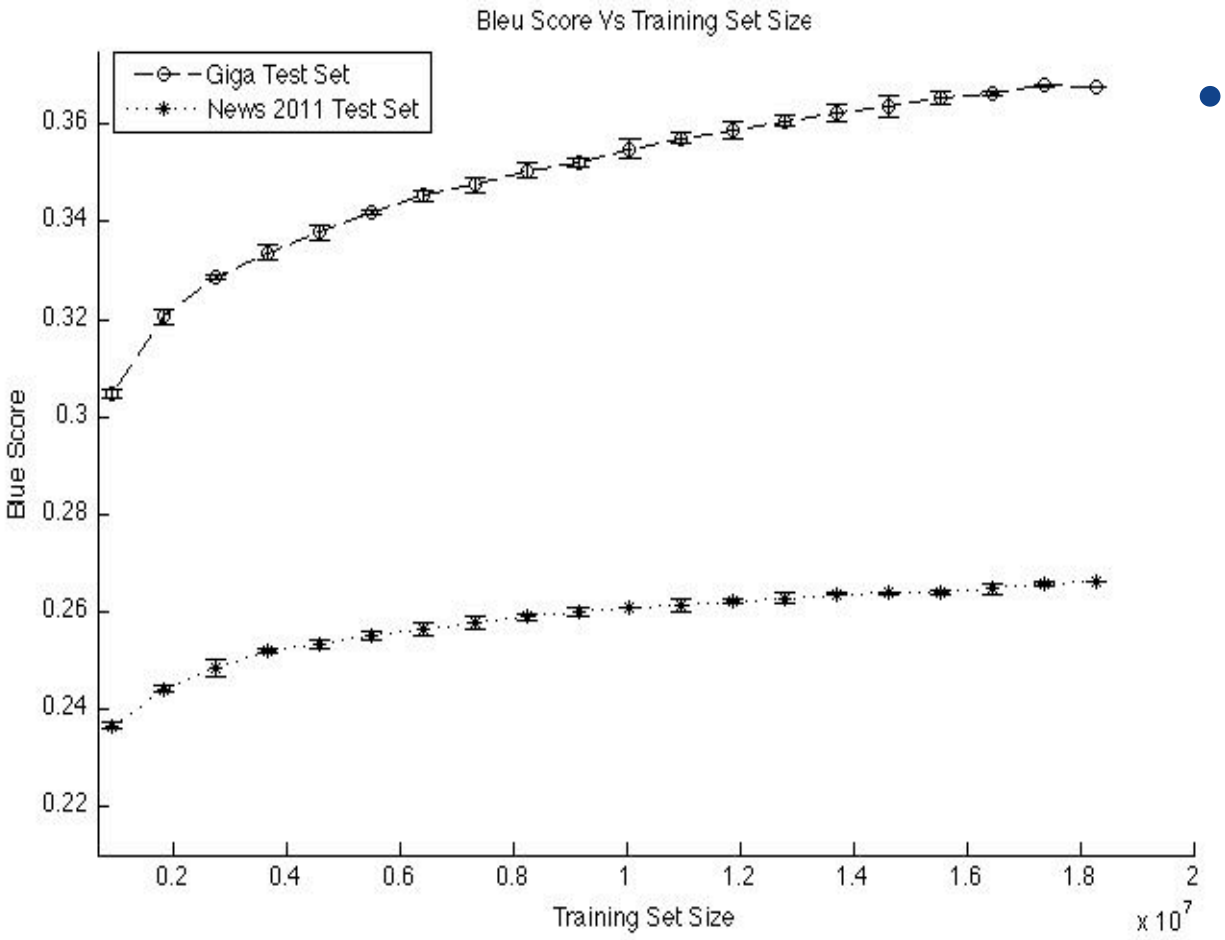
- Changing test set from the same distribution:
  - Same shape;
  - Small variation in the absolute Bleu score values.
- Considerations:
  - **There are not so huge corpora**;
  - Adding huge amount of data may affect **translation time**.
- **Freely available corpora** (when it is possible):
  - JRC-Acquis, News Commentaries, Europarl, UN Corpora, Movies Subtitle, MultiUN, Opus,...
- Some numbers:
  - Fa ~0.7M
  - De, It, Fr, Es, Pl ~ 4M
  - Ar ~ 6M
  - Tr, Pl > 16M using subtitle corpus.

JRC
EUROPEAN COMMISSION

Key Questions

- Five key questions when you build a translation service for news and not only:
  1. Which is the most suitable SMT system for our environment?
     - ***PBSMT***
  2. How to translate Named Entities in news?
     - ***suggests their correct translation to the SMT system***
  3. How to deal with different language styles in news?
     - ***Title and Content customization***
  4. How many data do we need?
     - ***All freely (and not) available corpora***
  5. **Is the news domain a problem**? (Work in Progress)

Is the news domain a problem?



- Testing on news data (out-of-domain):
  - **Drop in performance** compared to the in-domain test;
  - More **flat curve**;
  - **Small gain** in Bleu score.

- Everyday, **news topic changes** according to world events.
  - Always different and new words/phrases to translate.

JRC
EUROPEAN COMMISSION

Is the news domain a problem?

- News domain is dynamic.
- Parallel corpora
  - Are **static**;
  - Most of them are **not in the news domain**.
- Several approaches have been developed to gather new parallel sentences or fragments.
- Monolingual data varies:
  - daily **update of the language model** using English news of the day (Working in progress).

JRC
EUROPEAN COMMISSION

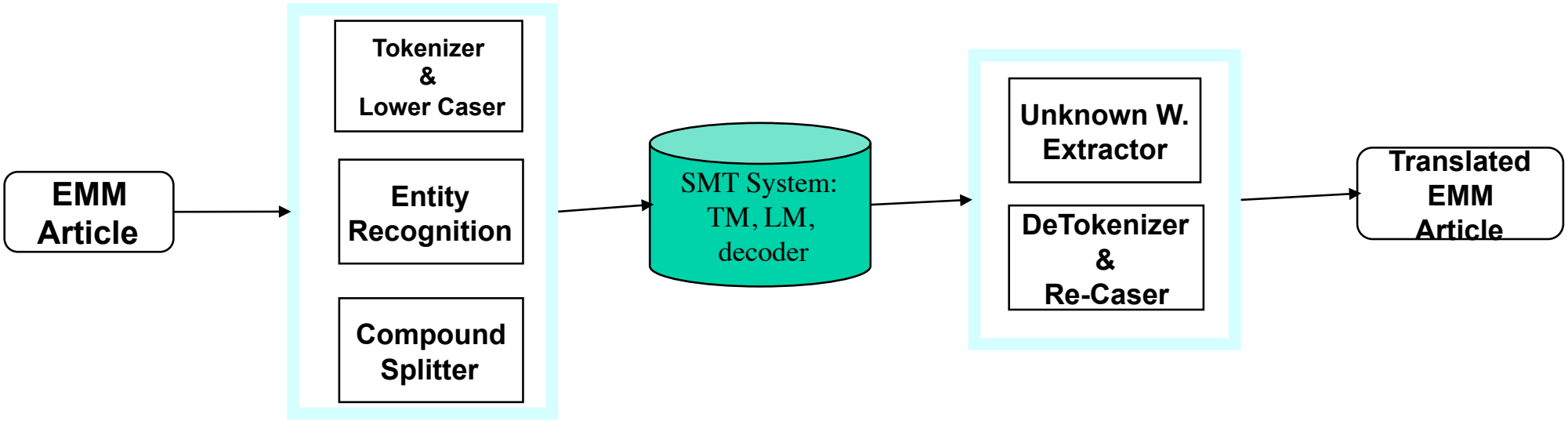
Key Questions

- Five key questions when you build a translation service for news and not only:
  1. Which is the most suitable SMT system for our environment?
     - ***PBSMT***
  2. How to translate Named Entities in news?
     - ***Suggests their correct translation to the SMT system***
  3. How to deal with different language styles in news?
     - ***Title and Content customization***
  4. Which training data can we use?
     - ***All freely available corpora***
  5. Is domain of the news a problem? (Work in Progress)
     - ***Language model updates***

- **Demo** (Thanks to Česlav Przywara and Yashar Mehdad)

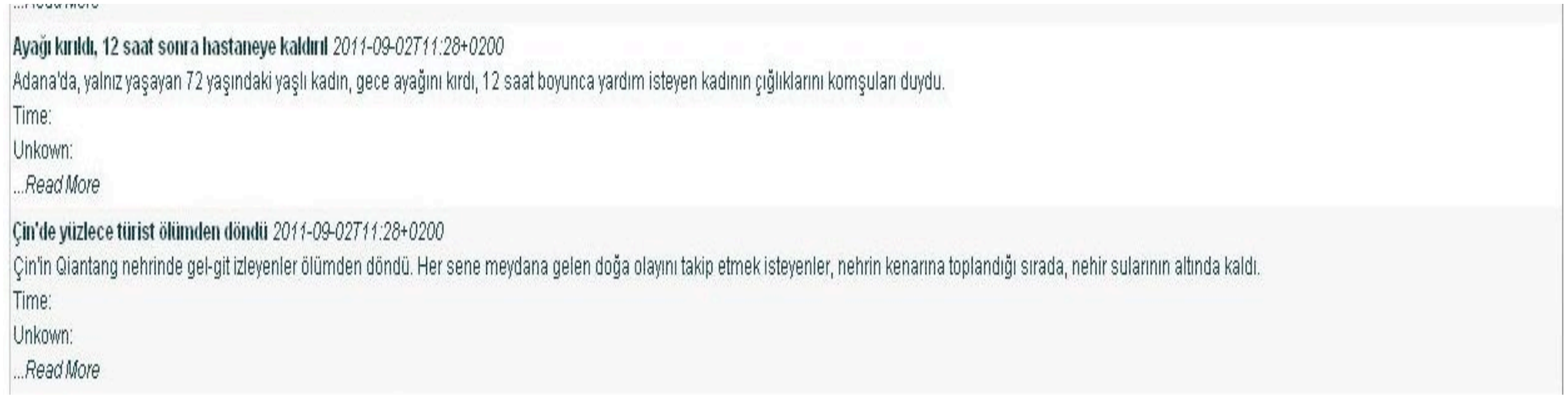
Demo

- Translation models available in **Fr, It, Es, De, Fa, Tr, Pl**.
  - **Ar** is coming soon!
- Based on **PBSMT** in **Moses**.
- EMM articles are previously categorized.

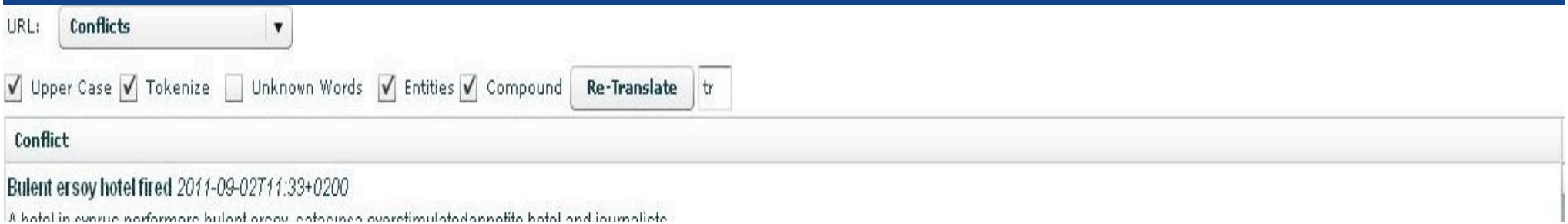


Demo – Tr → En

Libya'nın banknotları neden İngiltere'de basıldı ?
2011-09-02T11:24+0200
Kaddafi'nin devrilmesi ardından İngiliz uçaklarının ülkeye taşıdığı milyonlarca dinar değerindeki gıcır gıcır banknotun öyküsü.
Time:

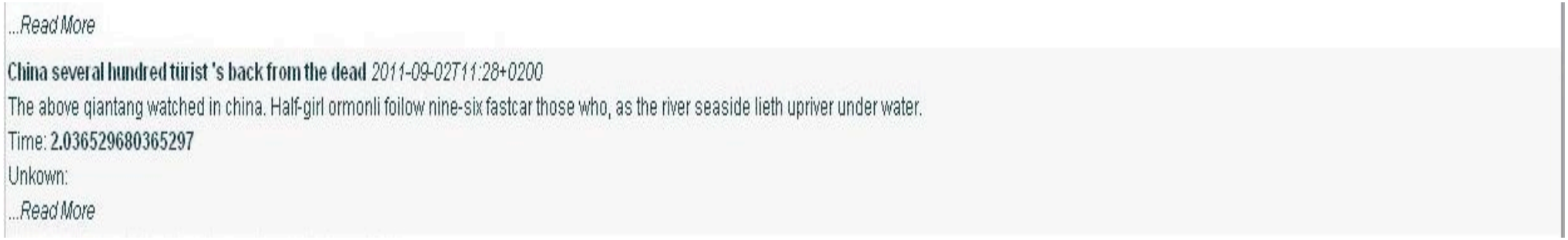


Demo – Tr → En



Libya'nın banknotları neden İngiltere'de basıldı ?
2011-09-02T11:24+0200
Kaddafi'nin devrilmesi ardından İngiliz uçaklarının ülkeye taşıdığı milyonlarca dinar değerindeki gıcır gıcır banknotun öyküsü.
Time:

Libya banknotes why a f[redacted] eject you know, he could be doing this?
2011-09-02T11:24+0200
This british aircraft lovingest Kadhafi toppling the millions of denars worth crisp banknotun story.
Time: 2.1509



Demo – Fa → En

CommunicableDiseases

عفونت مانع شایع تری عفونت در زنان 2011-07-06T10:27+0200

بیماری از میکروب ها از دستگاه گوارش وارد مجاری ادراری می شوند و موجب عفونت مثانه می شوند که داریم میان انجمل به باکتری ای - کالی (E.COLI) شایع تری عامل عفونت مثانه در زنان است.

Time:

Unkown:

...Read More

عفونت «E.coli» بیش از درمان، عوارض بلندمدت کلیوی دارد 2011-06-19T10:42+0200

سلامت نیوز : در حالی که چند روز پس از عفونت «E.coli» در اروپا به کند بیچ آرام می‌گیرد اما برخی از کارشناسان نسبت به پیامدها و خطرات بلندمدت انجمل به این عفونت در کلیه‌های هشدار دادند. به گزارش ایسنا، این کارشناسان اعلام کردند کسانی که به این

چه غذاهائی برای زنان باردار مضر است ؟

زنان باردار باید از مواد غذائی غیر پاستوریزه پرهیز کنند، زیرا موجب آسیب جنین و همچنین مادر میشود و اختلالاتی از جمله آسیب دیدن جفت ، عفونتهای خونی و یا مسموم شدن جنین و یا حتا سقط جنین را منجر میشود

Time:

سلامت نیوز : مقامات آلمانی می‌گویند به احتمال زیاد منشا شیوع «E.coli» که تاکنون 22 نفر را به هلاکت رسانده است نوعی جوانه لوبیای است. به گزارش ایسنا ، ایالات ساکسونی سفلی در شمال آلمان به شهروندان هشدار داد تا از خوردن جوانه لوبیای به دلیل این که احتمال میرود منشا شیوع E. coli باشد اجتناب ورزند.

Time:

Unkown:

...Read More

چه غذاهایی برای زنان باردار مضر است؟ 2011-07-16T16:06+0200

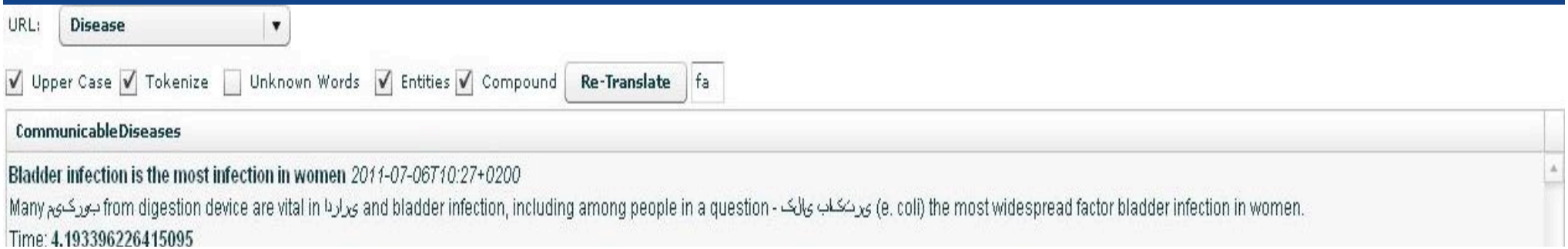
سلامت نیوز : غذاهای زیادی وجود دارد که برای سلامتی زنان باردار مفید است و در مقابل آن، غذاهایی نیز وجود دارد که عوارض برای اینکه سلامت مادر را تهدید می‌کنند، به همین نیز خطراتی را برای می‌کنند. زنان باردار باید از مواد غذایی غیر پاستوریزه پرهیز کنند، نیز اجتناب آسیب جنین و همچنین مادر می‌شوند و اختلالاتی....

Time:

Unkown:

...Read More

Demo – Fa → En



چه غذاهائی برای زنان باردار مضر است ؟

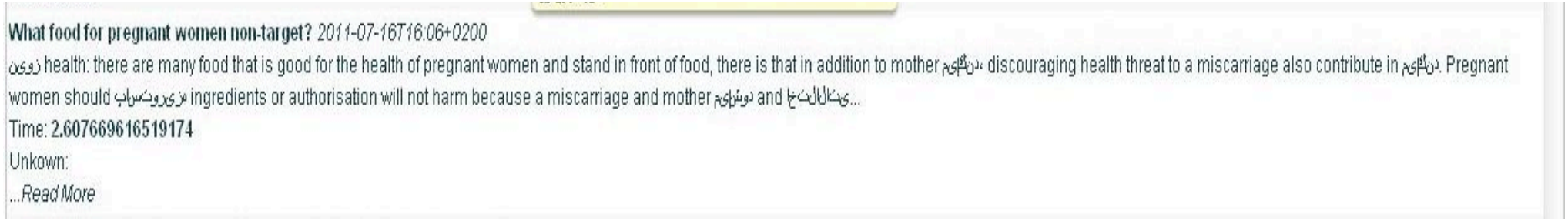
زنان باردار باید از مواد غذائی غیر پاستوریزه پرهیز کنند، زیرا موجب آسیب جنین و همچنین مادر میشود و اختلالاتی از جمله آسیب دیدن جفت ، عفونتهای خونی و یا مسموم شدن جنین و یا حتا سقط جنین را منجر میشود

Time:

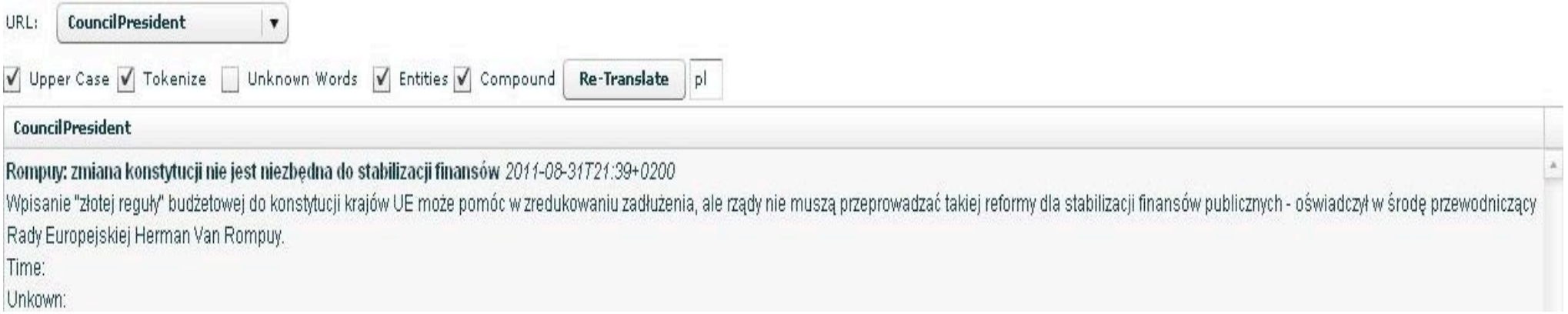
What *UNK* for pregnant women non-target?

Pregnant women should avoid sterile authorisation provisions because cause damage to the fetus and also *UNK* mother and *UNK* including harm to see *UNK UNK* blood or loco fetus or even the eruption of a spontaneous uprising, an abortion.

Time: 2.361



Demo – Pl → En



Sikorski o sytuacji w Libii: dziś jest dzień satysfakcji
2011-08-23T02:41+0200

Prezydencja Szef polskiej dyplomacji Radosław Sikorski wyraził nadzieję, że Libijczycy powrócą do społeczności międzynarodowej, z demokratycznie wybranymi władzami. W czasie poniedziałkowego briefingu poświęconego sytuacji w Libii, nie chciał natomiast komentować doniesień o tym, że Polska dostarczała broń libijskim powstańcom.

Time:

Demo – Pl → En

Sikorski o sytuacji w Libii: dziś jest dzień satysfakcji
2011-08-23T02:41+0200

Prezydencja Szef polskiej dyplomacji Radosław Sikorski wyraził nadzieję, że Libijczycy powrócą do społeczności międzynarodowej, z demokratycznie wybranymi władzami. W czasie poniedziałkowego briefingu poświęconego sytuacji w Libii, nie chciał natomiast komentować doniesień o tym, że Polska dostarczała broń libijskim powstańcom.

Time:

Read More

Sikorski on the situation in Libya: today is the day of satisfaction
2011-08-23T02:41+0200

The head of the Polish diplomacy Radoslaw Sikorski expressed the hope that *libijczycy* will return to the international community of democratically elected authorities. On Monday during a briefing on the situation in Libya, however, did not want to comment on reports that Poland *powstańcom* Libyan supplied weapons.

Time: 1.55

JRC
EUROPEAN COMMISSION

Open Issues

- **Word ordering** in the target sentence.
- **Vocabulary coverage** in agglutinative language, e.g. Tr.
- Select the **most reliable translated articles** to show to the final user
  - Confidence estimation approaches.
- Training data for **less resourced languages** or approaches to **expand translation capability without parallel data**
  - Lithuanian, Farsi, Croatian, Swahili,…
- **Subtitle corpora**
  - Pl-En 4M training sentences without subtitles: 57.84 Bleu score;
  - Pl-En 19M training sentences with subtitles: 57.36 Bleu score;
  - Tr-En with subtitle: high frequency of slang words.
- Different perception of translation quality (ready to luggage pack)
- …

JRC
EUROPEAN COMMISSION

Conclusions

- Large amount of data in more than 50 languages are gathered by the EMM's family every day.

- Language-independent algorithms:
  - Quite a **successful story at JRC**:
    - Multilingual named entity recognition and variant mapping
    - …

- Machine Translation is a **new research area** at the JRC.
  - Translation service based on Moses:
    - Translate documents to a common language.

  - Use the translation engine to improve language independent algorithms.

Multilingual text mining and Machine Translation activities carried out at the EC's Joint Research Centre (JRC)

Marco Turchi
& the JRC's *OPTIMA* team – Open Source Text Information Mining and Analysis

Technical details and publications: http://langtech.jrc.ec.europa.eu/
Applications: http://emm.newsbrief.eu/overview.html