Université de Montréal

Natural Language Processing on the Web

Guy Lapalme
RALI-DIRO, Université de Montréal

http://www.iro.umontreal.ca/~lapalme

rali

Overview

- What is Natural Language Processing (NLP)
- NLP for the Web
- The Web for NLP

rali

What is NLP ?

- Facilitates interactions between computers and humans
- Combines
  - Computer Science
  - Linguistics
  - Cognitive Science

rali

NLP approaches

- Rationalist (1950-…)
  - human is born with a language proficiency
  - symbolic approach
- Empiricist (1980-…)
  - human is born with proficiencies in pattern recognition, inference and generalization
  - statistical / machine learning approach

rali

Classical applications of NLP

- Natural language parsing
- Natural language generation
- Spelling and grammar checker
- Automatic summarization
- Machine translation
- Information retrieval
- Speech recognition

rali

http://rali.iro.umontreal.ca

rali

NLP for the *syntactic* Web
search engines

- NLP saved the Web!
- Indexing of billions of pages
- Query analysis

rali

Query

User Interface

Caching

Ranking

Query-time computing

Inverted Index

Cached Pages

Page & Site Statistics

Index Builder

Web Page Parser

Pages

Links & Anchors

Crawler

Web

Link Map

Page Authority

Link Graph Builder

Link Analysis

Link Graph

Offline computing

Source: Jian-Yun Nie

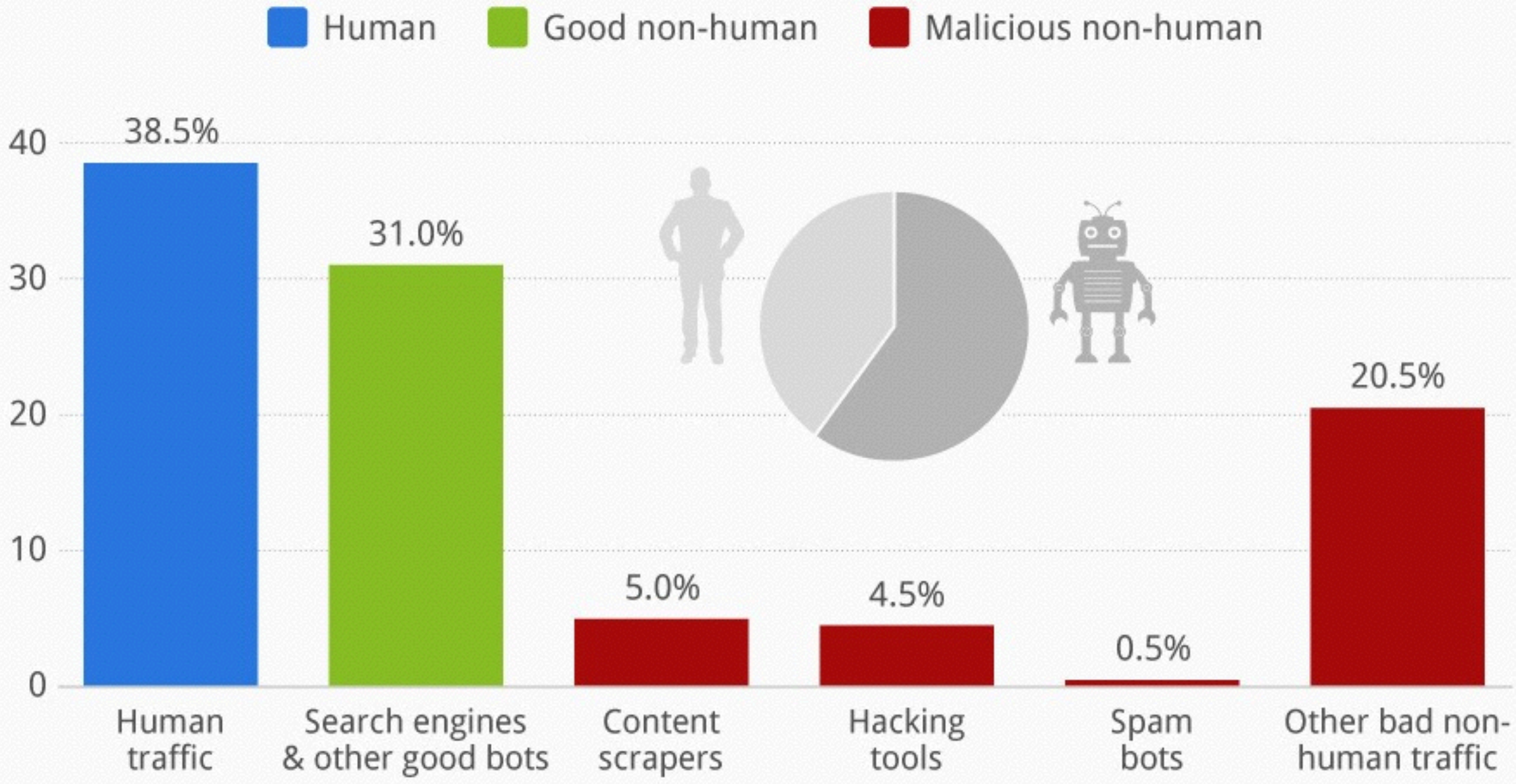
NLP is essential for the Web

- Most of the information is in human language
- But most *users* are computers !

rali

Humans Account for Less Than 40% of Global Web Traffic

Breakdown of global website traffic by source* (2013)

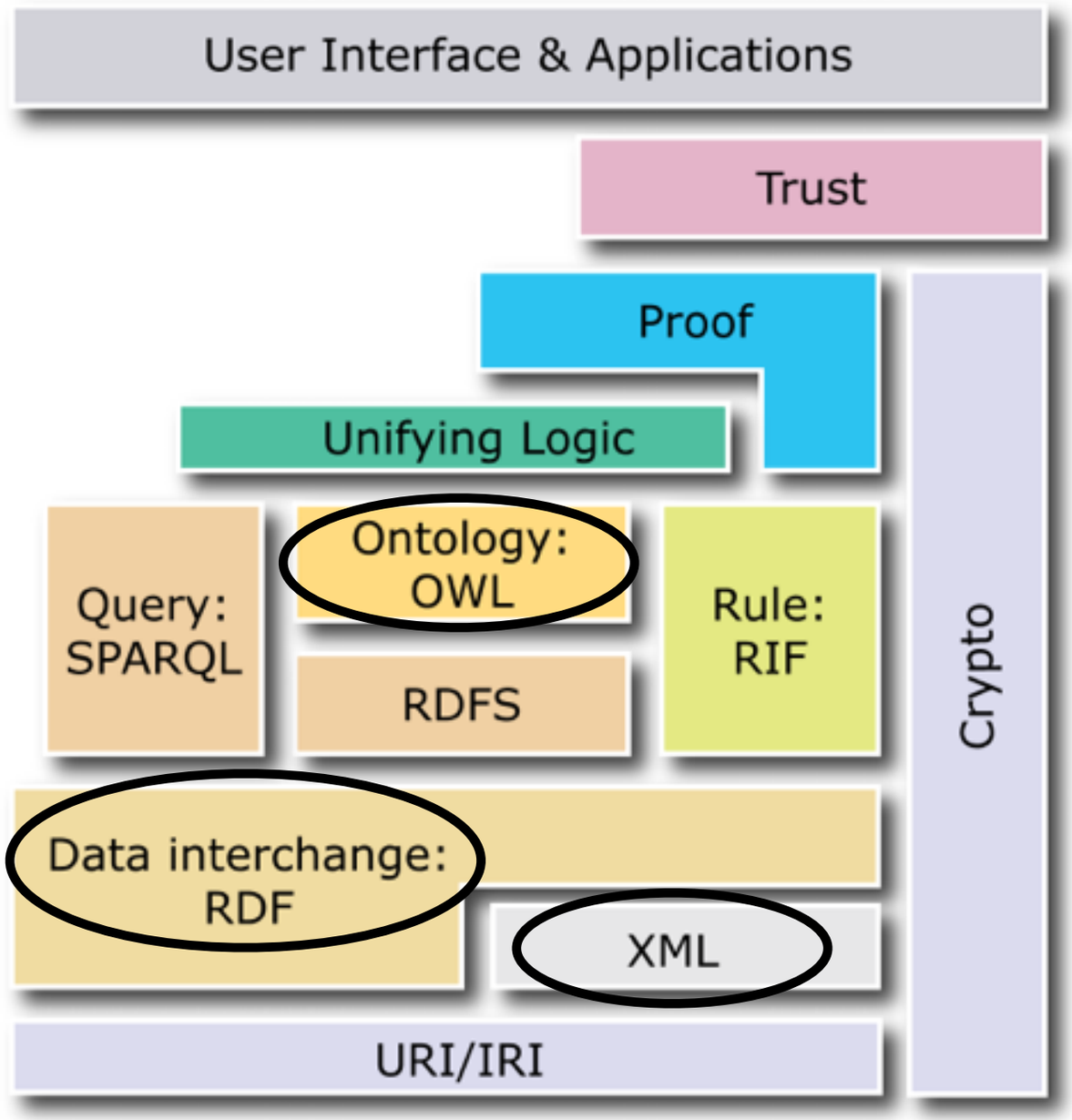


* based on 1.45 billion visits on 20,000 websites from 249 countries

Source: Incapsula

Mashable statista

NLP for the *semantic* Web



Humans

- want to continue writing and speaking as they ordinarily do
- are reluctant to type RDF or OWL

rali

NLP for the *semantic* Web
annotation challenge - I

- XML/HTML tags information
  - ideally done by the information provider
  - often limited to presentation info `<b>...</b>`
  - useless for application that needs content info

rali

HTML page as seen by a Machine or a Human

rali

NLP for the *semantic* Web *annotation challenge - 2*

- NLP through information extraction (IE) can help find and tag semantic info
  - named entity recognizer
  - date recognizer

rali

NLP for the *semantic* Web *annotation challenge - 3*

- RDF
  - links a *subject (S)* and an *object (O)* through a *property (P)*
  - S, P and O must be located in the text
  - Identical S, P and O must have the same URI
- NLP for knowledge base population (KBP 2014)
  - word sense disambiguation
  - event detection
  - sentiment analysis
  - entity linking

rali

NLP for the *semantic* Web *annotation challenge - 4*

- Ontologies (OWL)
  - individuals (*John, Mary, this table*)
  - classes (*man, woman, furniture*)
  - properties (*marriedTo, age, builtBy*)
- NLP for
  - populating an existing ontology (*OwlExporter*)
  - creating an ontology
    - seed concepts to collect entities
    - collect dependencies (identified by a parser) between entities
    - extract properties from path between concepts

rali

NLP for the Web

- Essential
- Not often recognized as such
- Web has given a strong push to NLP in recent years

rali

The Web for NLP

- Access to a practically infinite corpus of text in many languages
- Interesting facts about language given by collaborative sites
  - Wikipedia
  - DBpedia
  - Wiktionary

Language models

- Probability distribution on the strings of a language
- Computed by counting frequencies of words/ characters over texts (the more the better)
  - Google Web Trillion Word Corpus
  - Canadian Hansard (Bilingual: English-French)
  - Europarl (Multilingual)

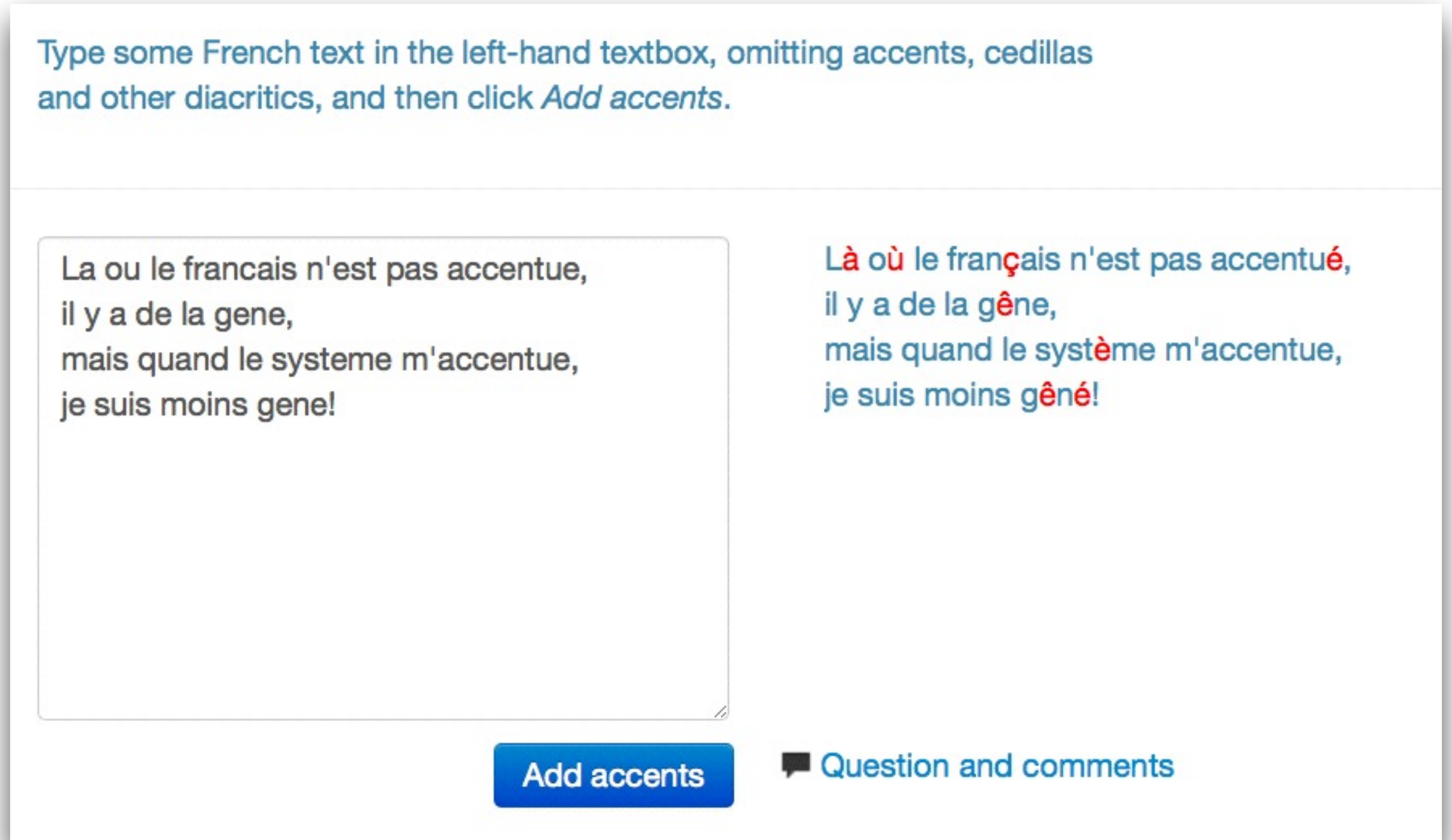
rali

Uses of unilingual language models

- Document classification
  - compare a text to the LM of a type of text
- Information retrieval
  - how does the LM of query match the one of a document
- Language identification
  - how does a sequence of characters compare with *usual* sequences in a given language
- Automatic accentuation
  - which possible accentuation best matches a *French* sentence

rali

http://rali.iro.umontreal.ca/rali/zodiac-project



One error on average every 200 words

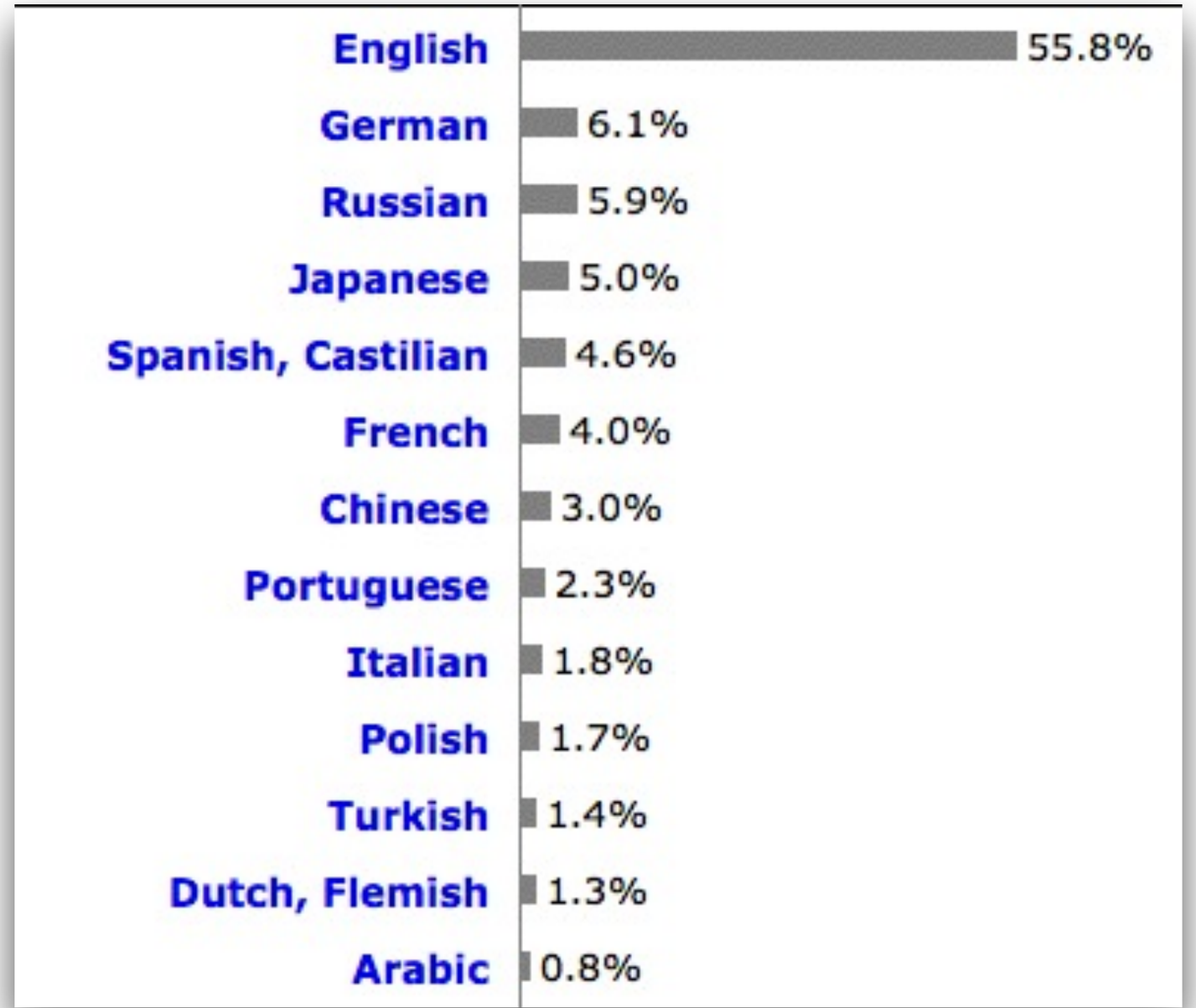
rali

Uses of bilingual texts on the web

- Bilingual concordancer
  - search engine in data base of sentences and their corresponding translations

Languages on the Web

English is used by 56% of all web sites



Source: w3techs.com

rali

TRANSSEARCH TERMINOTIX rali

USER : **lapalme** QUERIES | MY ACCOUNT | PREFERENCES | HELP | QUIT

Bilingual query

Document collection : House of Commons Hansard (1986-2012)

Expression: kettle of fish Search

13 translations of ***kettle of fish*** within 27 occurrences

paire de manches	14
situation bien	2
boîte de pandore	1
passe aux vraies affaires	1
problème	1
éloigner du poisson	1
dans de beaux draps	1
panier de crabes	1
bien différente	1
aux privilèges	1
histoire	1
changent la donne	1
chaudronnée de boullabaisse	1

paire de manches 14

If the hon. member has not been to look at them, that is another **kettle of fish**.	S'il n'est pas allé voir, c'est une autre **paire de manches**.
That would be an amazing proposal, but it is a different **kettle of fish** and not what we are working on today.	Bien que ce soit un beau grand projet, c'est une autre **paire de manches** et ça ne nous concerne pas aujourd'hui.
If no one stands up and everyone is happy with the apology, that is another **kettle of fish**.	Lorsque personne n'intervient et que la Chambre accepte les excuses, c'est une autre **paire de manches**.
He may be right about 1980, but 1995 was another **kettle of fish**.	Effectivement, en 1980, cela pouvait s'appliquer, mais en 1995, c'était une autre **paire de manches**.
It was a very different **kettle of fish**.	C'était une autre **paire de manches**.
If the members opposite do not trust judges, that is a different **kettle of fish**.	S'ils ne font pas confiance aux juges, c'est une autre **paire de manches**.
This is a different **kettle of fish**, because a conservative generally opposes this kind of spending.	Là, nous avons une nouvelle **paire de manches**, car si vous êtes conservateurs, vous êtes contre ce genre de dépenses.
However, when they won the elections, it was a different **kettle of fish**.	Toutefois, lorsqu'ils ont remporté les élections, c'était alors une toute autre **paire de manches**.
It is a different **kettle of fish**.	C'est une autre **paire de manches**.
Question period is a different **kettle of fish**.	La période des questions, c'est une autre **paire de manches**.

1 / 2 Pages

rali tsrali.com

Uses of bilingual texts on the web

- Machine translation

$$\text{argmax}_{e} P(e \mid f) = \text{argmax}_{e} P(f \mid e)\, P(e)$$

  - $P(f \mid e)$: transfer model between source and target computed from aligned bilingual texts
  - $P(e)$: language model for the target language
  - decoder to find the best solution

rali

phrase based model

```
have not been democratically |||
          n' ont été démocratiquement ||| 0.5 9.81223e-08
particularly in baby products . |||
         en particulier dans les produits pour bébés . ||| 0.5 4.17267e-12
there is no cruelty |||
          il n' est nullement question de cruauté ||| 1 3.15409e-10
intergovernmental conference and reform of the treaty |||
          cig et la révision du traité ||| 1 1.03844e-17
is an absolute must . |||
          c' est un nécessité absolue . ||| 1 1.35065e-12
parliament has already done a considerable amount |||
          le parlement a accompli un travail énorme ||| 1 8.38686e-21
presidency 's programme |||
          du programme de la présidence ||| 1 9.33702e-07
```

rali

Source: Philippe Langlais

http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html

WATT
Warning-Avertissement Translation-Traduction

En français

rali

This prototype has been developed to translate Public Weather Warnings published daily by Environment Canada.

Enter some text in the left text area and press Translate, or click to get sample source text in French or English

Source text

ON PREVOIT DES VENTS FORTS D OUEST EN RAFALES A 90 KM/H OU PLUS SUR LA MAJEURE PARTIE DU SUD DE L ALBERTA. DES VENTS SE LEVERONT SUR LES CONTREFORTS ET SE PROPAGERONT VERS L EST AU PASSAGE DU FRONT.

Des vents violents du secteur nord sont attendus cette nuit sur Blanc-Sablon et ils persisteront jusqu'à samedi matin.

À 21h30, la neige avait faibli peu à peu à Gillam et Shamattawa, mais elle se maintenait à Tadoule Lake et Brochet. Cependant, les vents d'est faibliront cette nuit, puis des vents du nord très forts de 60 km/h soufflant en rafales jusqu'à 80 km/h se lèveront samedi après-midi et se maintiendront pendant la nuit, occasionnant de la poudrerie haute et une visibilité nulle.

Translation

Strong westerly winds gusting to 90 km/h or greater are expected throughout much of southern Alberta. Winds will develop over the foothills spreading eastwards with the frontal passage.

High northerly winds are expected tonight over Blanc-Sablon and they will persist through Saturday morning.

At 9:30 PM, the snow had tapered off at Gillam and Shamattawa, but continued at Tadoule Lake and Brochet. However, easterly winds will diminish tonight, then very strong north winds of 60 km/h gusting to 80 km/h will develop Saturday afternoon and continue through the night, giving zero visibility in blowing snow.

Translation direction: Autodetected

Translate »

rali

Web for NLP for the Web

- Machine translation of Web pages
  - Google Translate
- Ontologies for helping users
  - Yellow pages
- Speech recognition
  - Siri

the WEB

would not be the same without

NLP

NLP

would not be the same without

the WEB

References

- Y. Wilks, C. Brewster, Natural Language Processing as a Foundation of the Semantic Web, *Web Science*, vol 1, nos 3-4, pp 199-327, 2006.
- Knowledge Base Population (KBP) 2014 - TAC 2014 Track
- B. Liu, Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data, 2nd ed., Springer, 2011, 532 p.
- Guy Lapalme, XML: Looking at the Forest Instead of the Trees.
- Lapalme, G., P. Langlais, F. Gotti, The Bilingual Concordancer TransSearch, *NAACL 2012 - Demo-Poster*.
- Gotti, F., P. Langlais, G. Lapalme, Designing a Machine Translation System for Canadian Weather Warnings: a Case Study, *Natural Language Engineering*, vol. 20, issue 3: Cambridge University Press, pp. 399-433, 07/2014.

rali