A Factory of Comparable Corpora from Wikipedia

Alberto Barrón-Cedeño[1], Cristina España-Bonet[2],
Josu Boldoba[2], and Lluís Màrquez[1]

[1] Qatar Computing Research Institute, HBKU, Qatar
[2] TALP Research Center, UPC, Spain
{albarron, lmarquez}@qf.org.qa
cristinae@cs.upc.edu jboldoba08@gmail.com

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

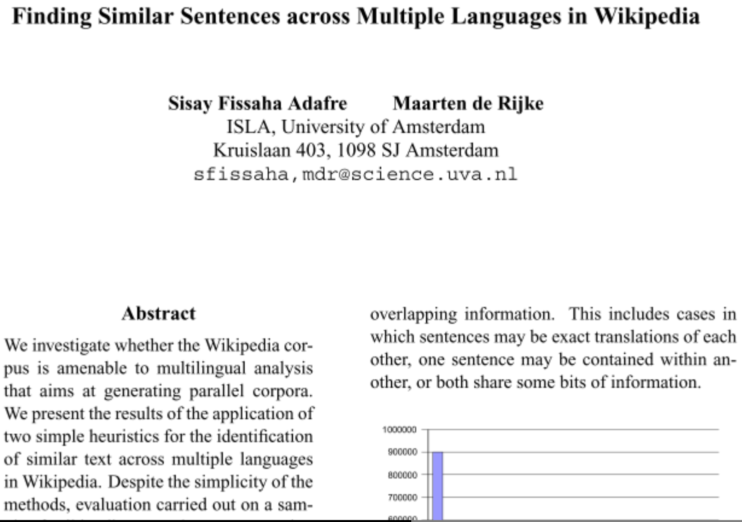
عضو في مؤسسة قطر
Member of Qatar Foundation

UPC

BUCC @ ACL – July 2015

Background

There are tons of articles exploiting Wikipedia as a comparable corpus



Finding Similar Sentences across Multiple Languages in Wikipedia

Sisay Fissaha Adafre **Maarten de Rijke**
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
sfissaha,mdr@science.uva.nl

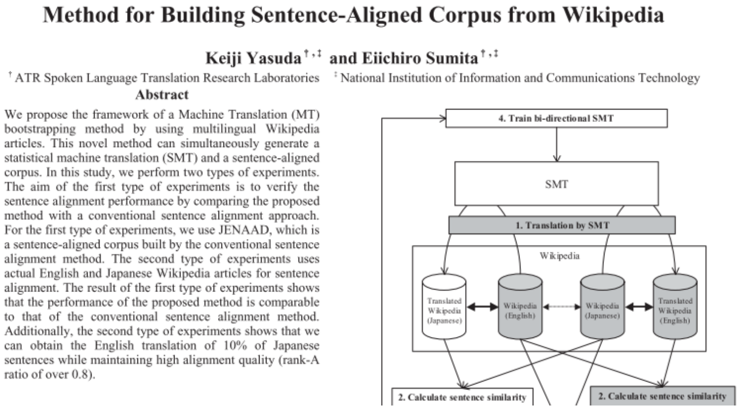
Abstract

We investigate whether the Wikipedia corpus is amenable to multilingual analysis that aims at generating parallel corpora. We present the results of the application of two simple heuristics for the identification of similar text across multiple languages in Wikipedia. Despite the simplicity of the methods, evaluation carried out on a sam-

overlapping information. This includes cases in which sentences may be exact translations of each other, one sentence may be contained within another, or both share some bits of information.

Background

There are tons of articles exploiting Wikipedia as a comparable corpus



Method for Building Sentence-Aligned Corpus from Wikipedia

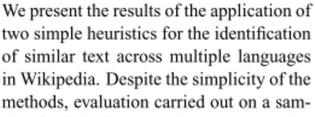
Keiji Yasuda[†,‡] and Eiichiro Sumita[†,‡]

[†] ATR Spoken Language Translation Research Laboratories [‡] National Institution of Information and Communications Technology

Abstract

We propose the framework of a Machine Translation (MT) bootstrapping method by using multilingual Wikipedia articles. This novel method can simultaneously generate a statistical machine translation (SMT) and a sentence-aligned corpus. In this study, we perform two types of experiments. The aim of the first type of experiments is to verify the sentence alignment performance by comparing the proposed method with a conventional sentence alignment approach. For the first type of experiments, we use JENAAD, which is a sentence-aligned corpus built by the conventional sentence alignment method. The second type of experiments uses actual English and Japanese Wikipedia articles for sentence alignment. The result of the first type of experiments shows that the performance of the proposed method is comparable to that of the conventional sentence alignment method. Additionally, the second type of experiments shows that we can obtain the English translation of 10% of Japanese sentences while maintaining high alignment quality (rank-A ratio of over 0.8).

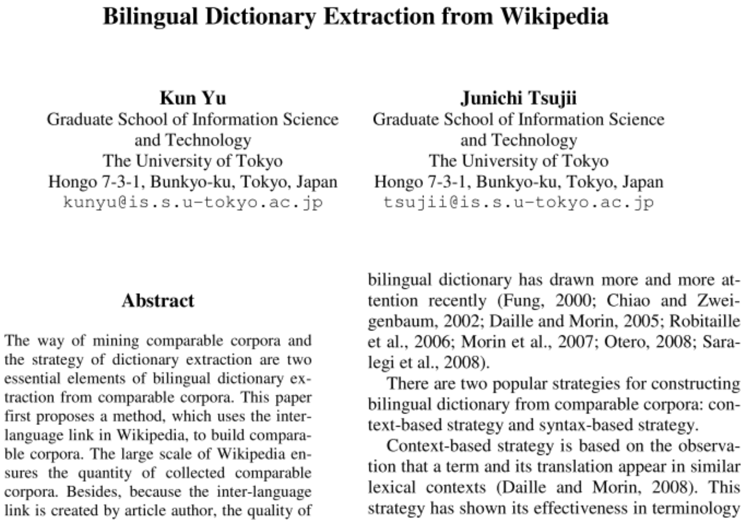
4. Train bi-directional SMT
SMT
1. Translation by SMT
Wikipedia
Translated Wikipedia (Japanese)
Wikipedia (English)
Wikipedia (Japanese)
Translated Wikipedia (English)
2. Calculate sentence similarity
2. Calculate sentence similarity



We present the results of the application of two simple heuristics for the identification of similar text across multiple languages in Wikipedia. Despite the simplicity of the methods, evaluation carried out on a sam-

Background

There are tons of articles exploiting Wikipedia as a comparable corpus



Bilingual Dictionary Extraction from Wikipedia

Kun Yu
Graduate School of Information Science
and Technology
The University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan
kunyu@is.s.u-tokyo.ac.jp

Junichi Tsujii
Graduate School of Information Science
and Technology
The University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan
tsujii@is.s.u-tokyo.ac.jp

Abstract

The way of mining comparable corpora and the strategy of dictionary extraction are two essential elements of bilingual dictionary extraction from comparable corpora. This paper first proposes a method, which uses the inter-language link in Wikipedia, to build comparable corpora. The large scale of Wikipedia ensures the quantity of collected comparable corpora. Besides, because the inter-language link is created by article author, the quality of

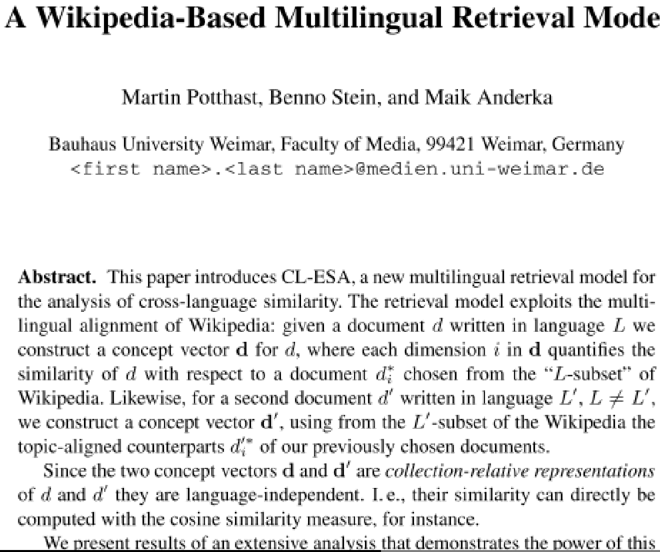
bilingual dictionary has drawn more and more attention recently (Fung, 2000; Chiao and Zweigenbaum, 2002; Daille and Morin, 2005; Robitaille et al., 2006; Morin et al., 2007; Otero, 2008; Saralegi et al., 2008).

There are two popular strategies for constructing bilingual dictionary from comparable corpora: context-based strategy and syntax-based strategy.

Context-based strategy is based on the observation that a term and its translation appear in similar lexical contexts (Daille and Morin, 2008). This strategy has shown its effectiveness in terminology

Background

There are tons of articles exploiting Wikipedia as a comparable corpus



A Wikipedia-Based Multilingual Retrieval Model

Martin Potthast, Benno Stein, and Maik Anderka

Bauhaus University Weimar, Faculty of Media, 99421 Weimar, Germany
<first name>.<last name>@medien.uni-weimar.de

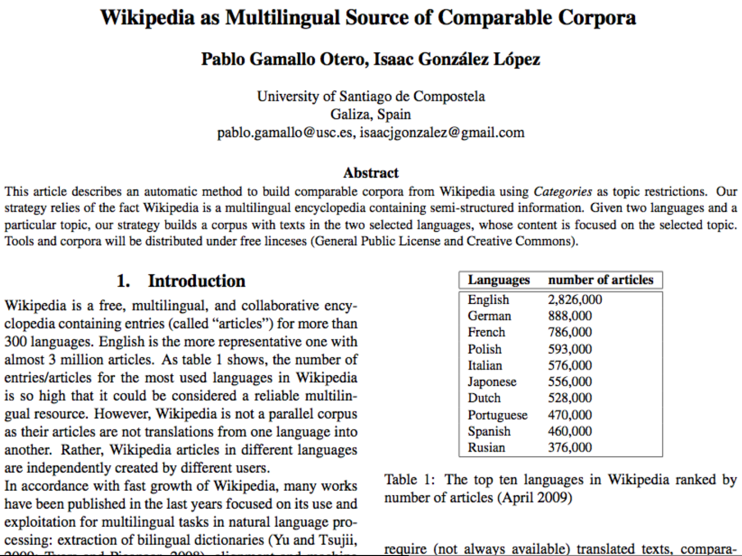
Abstract. This paper introduces CL-ESA, a new multilingual retrieval model for the analysis of cross-language similarity. The retrieval model exploits the multilingual alignment of Wikipedia: given a document d written in language L we construct a concept vector $\mathbf{d}$ for d, where each dimension i in $\mathbf{d}$ quantifies the similarity of d with respect to a document d_i^* chosen from the "L-subset" of Wikipedia. Likewise, for a second document d' written in language L', $L \neq L'$, we construct a concept vector $\mathbf{d}'$, using from the L'-subset of the Wikipedia the topic-aligned counterparts $d_i'^*$ of our previously chosen documents.

Since the two concept vectors $\mathbf{d}$ and $\mathbf{d}'$ are *collection-relative representations* of d and d' they are language-independent. I. e., their similarity can directly be computed with the cosine similarity measure, for instance.

We present results of an extensive analysis that demonstrates the power of this

Background

There are tons of articles exploiting Wikipedia as a comparable corpus



Wikipedia as Multilingual Source of Comparable Corpora

Pablo Gamallo Otero, Isaac González López

University of Santiago de Compostela
Galiza, Spain
pablo.gamallo@usc.es, isaacjgonzalez@gmail.com

Abstract

This article describes an automatic method to build comparable corpora from Wikipedia using *Categories* as topic restrictions. Our strategy relies of the fact Wikipedia is a multilingual encyclopedia containing semi-structured information. Given two languages and a particular topic, our strategy builds a corpus with texts in the two selected languages, whose content is focused on the selected topic. Tools and corpora will be distributed under free linceses (General Public License and Creative Commons).

1. Introduction

Wikipedia is a free, multilingual, and collaborative encyclopedia containing entries (called "articles") for more than 300 languages. English is the more representative one with almost 3 million articles. As table 1 shows, the number of entries/articles for the most used languages in Wikipedia is so high that it could be considered a reliable multilingual resource. However, Wikipedia is not a parallel corpus as their articles are not translations from one language into another. Rather, Wikipedia articles in different languages are independently created by different users.

In accordance with fast growth of Wikipedia, many works have been published in the last years focused on its use and exploitation for multilingual tasks in natural language processing: extraction of bilingual dictionaries (Yu and Tsujii,

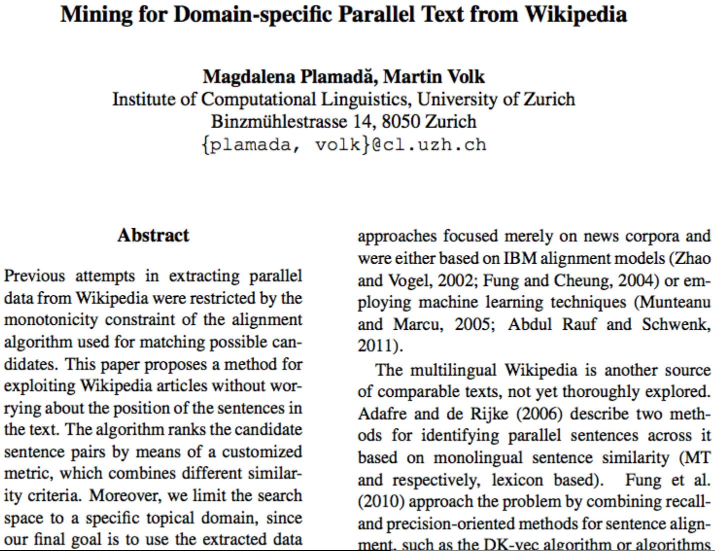
Languages	number of articles
English	2,826,000
German	888,000
French	786,000
Polish	593,000
Italian	576,000
Japonese	556,000
Dutch	528,000
Portuguese	470,000
Spanish	460,000
Rusian	376,000

Table 1: The top ten languages in Wikipedia ranked by number of articles (April 2009)

require (not always available) translated texts, compara-

Background

There are tons of articles exploiting Wikipedia as a comparable corpus



Mining for Domain-specific Parallel Text from Wikipedia

Magdalena Plamadă, Martin Volk
Institute of Computational Linguistics, University of Zurich
Binzmühlestrasse 14, 8050 Zurich
{plamada, volk}@cl.uzh.ch

Abstract

Previous attempts in extracting parallel data from Wikipedia were restricted by the monotonicity constraint of the alignment algorithm used for matching possible candidates. This paper proposes a method for exploiting Wikipedia articles without worrying about the position of the sentences in the text. The algorithm ranks the candidate sentence pairs by means of a customized metric, which combines different similarity criteria. Moreover, we limit the search space to a specific topical domain, since our final goal is to use the extracted data

approaches focused merely on news corpora and were either based on IBM alignment models (Zhao and Vogel, 2002; Fung and Cheung, 2004) or employing machine learning techniques (Munteanu and Marcu, 2005; Abdul Rauf and Schwenk, 2011).

The multilingual Wikipedia is another source of comparable texts, not yet thoroughly explored. Adafre and de Rijke (2006) describe two methods for identifying parallel sentences across it based on monolingual sentence similarity (MT and respectively, lexicon based). Fung et al. (2010) approach the problem by combining recall- and precision-oriented methods for sentence alignment, such as the DK-vec algorithm or algorithms

Background

Nevertheless…

- Little attention is paid to identifying a domain-specific high-quality comparable corpus
- Domain-specific corpora is a key factor in different tasks, including MT

Background

Nevertheless…

- Little attention is paid to identifying a domain-specific high-quality comparable corpus
- Domain-specific corpora is a key factor in different tasks, including MT
- Wikipedia includes (somehow) all the information necessary to extract such a resource

Background

Nevertheless…

- Little attention is paid to identifying a domain-specific high-quality comparable corpus
- Domain-specific corpora is a key factor in different tasks, including MT
- Wikipedia includes (somehow) all the information necessary to extract such a resource

Our aim is to identify those domain-specific comparable corpora from Wikipedia!

Background: Strategy Overview

- Identify comparable articles (easy)

Background: Strategy Overview

- Identify comparable articles (easy)
- Build a characteristic vocabulary for the domain of interest (not so easy)

Background: Strategy Overview

- Identify comparable articles (easy)
- Build a characteristic vocabulary for the domain of interest (not so easy)
- Explore the Wikipedia categories' graph to select the subset of categories in the domain (difficult)
- Brute-force sentence-wise comparison for parallel pairs identification

Outline

Background

Domain-Specific Comparable Corpora Extraction

Parallel Fragments Extraction

Impact on an MT System

Final Remarks

Comparable Corpora

Problem No large collections of comparable texts for all domains and language pairs exist

Objective To extract high-quality comparable corpora on specific domains

Pilot language pair English–Spanish

Pilot domains Science, Computer Science, Sports

Currently experimenting on more than 700 domains and 10 languages

Comparable Corpora: Characteristic Vocabulary

1. Retrieve every article associated to the top category of the domain (e.g., Sports)

Comparable Corpora: Characteristic Vocabulary

1. Retrieve every article associated to the top category of the domain (e.g., Sports)

2. Merge the articles' contents and apply standard and ad-hoc pre-processing

Comparable Corpora: Characteristic Vocabulary

1. Retrieve every article associated to the top category of the domain (e.g., Sports)

2. Merge the articles' contents and apply standard and ad-hoc pre-processing

3. Select the top-k tf-sorted tokens as the characteristic vocabulary

   (we consider 10% of the tokens)

Comparable Corpora: Characteristic Vocabulary

1. Retrieve every article associated to the top category of the domain (e.g., Sports)

2. Merge the articles' contents and apply standard and ad-hoc pre-processing

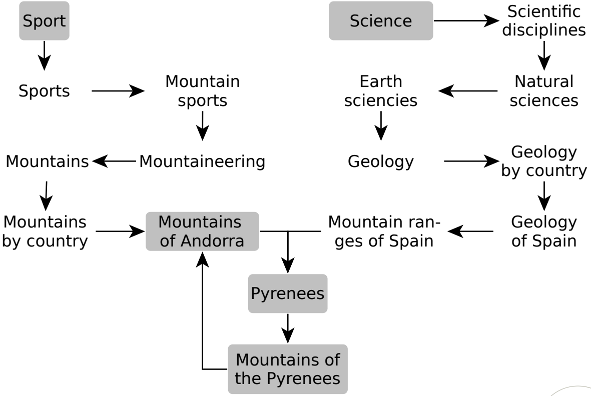
3. Select the top-k tf-sorted tokens as the characteristic vocabulary

   (we consider 10% of the tokens)

	Articles		Vocabulary	
	en	es	en	es
CS	4	130	106	447
Sc	29	3	464	140
Sp	3	10	122	100

Comparable Corpora: Graph exploration

Slice of the Spanish Wikipedia category graph departing from categories Sport and Science (as in Spring 2015)



Comparable Corpora: Graph exploration

1. Perform a breadth-first search departing from the root category
2. Visit nodes only once to avoid loops and repeating traversed paths
3. Stop at the level when most categories do not belong to the domain

Comparable Corpora: Graph exploration

1. Perform a breadth-first search departing from the root category
2. Visit nodes only once to avoid loops and repeating traversed paths
3. Stop at the level when most categories do not belong to the domain

   Stopping criterion

   Heuristic A category belongs to the domain if its title contains at least one term from the characteristic vocabulary

   Explore until a minimum percentage of the categories in a tree level belong to the domain

Comparable Corpora: Graph exploration

1. Perform a breadth-first search departing from the root category
2. Visit nodes only once to avoid loops and repeating traversed paths
3. Stop at the level when most categories do not belong to the domain

Stopping criterion

Heuristic A category belongs to the domain if its title contains at least one term from the characteristic vocabulary

Explore until a minimum percentage of the categories in a tree level belong to the domain

Category pato in Spanish —literally "duck"— refers to a sport rather than an animal!!!

Comparable Corpora: Graph exploration

Article pairs selected according to two criteria: 50% and 60%

	Articles		Distance from the root			
	50%	60%	50%		60%	
	en-es	en-es	en	es	en	es
CS	18,168	8,251	6	5	5	5
Sc	161,130	21,459	6	4	4	4
Sp	72,315	1,980	8	8	3	4

Outline

Background

Domain-Specific Comparable Corpora Extraction

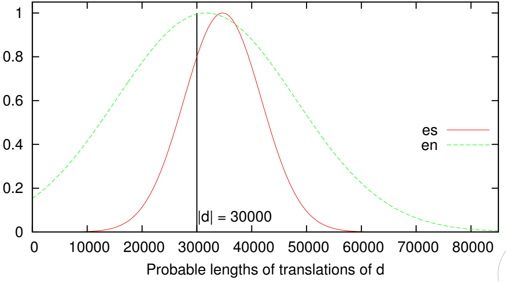
Parallel Fragments Extraction

Impact on an MT System

Final Remarks

Parallelisation: Similarity Models

- Character 3-grams (cosine) [McNamee and Mayfield, 2004]
- Pseudo-cognates (cosine) [Simard et al., 1992]
- Translated word 1-grams in both directions (cosine)
- Length factor [Pouliquen et al., 2003]



Parallelisation: Corpus for Preliminary Evaluation

- 30 article pairs (10 per domain)
- Annotated at sentence level
- Three classes: parallel, comparable, and other
- Each pair was annotated by 2 volunteers mean Cohen's $\kappa \sim 0.7$

Parallelisation: Threshold Definition

	c3g	cog	mono$_{en}$	mono$_{es}$	len
Thres.	0.25	0.30	0.20	0.15	0.90
P	0.28	0.16	0.30	0.26	0.08
R	0.53	0.49	0.46	0.34	0.57
F$_1$	0.36	0.24	0.36	0.30	0.14

Parallelisation: Threshold Definition

	c3g	cog	mono_{en}	mono_{es}	len
Thres.	0.25	0.30	0.20	0.15	0.90
P	0.28	0.16	0.30	0.26	0.08
R	0.53	0.49	0.46	0.34	0.57
F_1	0.36	0.24	0.36	0.30	0.14

	$\bar{S}$	$\bar{S}$·len	$\overline{S \cdot F_1}$	$\overline{S \cdot F_1}$·len
Thres.	0.25	0.15	0.05	0.05
P	0.27	0.33	0.18	0.32
R	0.50	0.62	0.77	0.65
F_1	0.35	0.43	0.29	0.43

S =similarities; $\bar{\cdot}$ = average

Parallelisation: Parallel Sentences

	CS	Sc	Sp
c3g	96,039	724,210	335,147
cog	182,981	1,215,008	451,941
len	271,073	1,941,866	550,338
mono_{en}	211,209	1,367,917	461,731
mono_{es}	183,439	1,273,509	435,671
$\bar{\text{S}}$	154,917	1,098,453	450,933
$\bar{\text{S}}\cdot\text{len}$	121,697	957,662	390,783
$\overline{\text{S} \cdot \text{F}_1}$	153,056	1,085,502	448,076
$\overline{\text{S} \cdot \text{F}_1}\cdot\text{len}$	121,407	957,967	392,241

Outline

Background

Domain-Specific Comparable Corpora Extraction

Parallel Fragments Extraction

Impact on an MT System

Final Remarks

Impact: Corpora

	in domain	out of domain
Training	Wikipedia	Europarl
Development	Wikipedia	News commentary
Test	Wikipedia/Gnome	News commentary

Impact: Corpora

	in domain	out of domain
Training	Wikipedia	Europarl
Development	Wikipedia	News commentary
Test	Wikipedia/Gnome	News commentary

Generation of the Wikipedia dev and test sets

1. Select only sentences starting with a letter and longer than three tokens
2. Compute the perplexity of each sentence pair (with respect to a Europarl LM)
3. Sort the pairs according to similarity and perplexity
4. Manually select the first k parallel sentences

Impact: Corpora Statistics

	CS	Sc	Sp	All
c3g	95,715	723,760	334,828	883,366
cog	182,283	1,213,965	451,324	1,430,962
mono$_{en}$	210,664	1,367,169	461,237	1,638,777
$\bar{S}$·len	120,835	956,346	389,975	1,160,977
union	577,428	3,847,381	1,181,664	4,948,241
Wikipedia dev	300	300	300	900
Wikipedia test	500	500	500	1500
Gnome	1000	–	–	–

Impact: Phrase-based SMT System

Language model 5-gram interpolated Kneser-Ney discounting, SRILM Toolkit

Alignments GIZA++ Toolkit

Translation model Moses package

Weights optimization MERT against BLEU

Decoder Moses

Impact: Experiments definition

1 In domain	Training	Wikipedia or Europarl
	Test	Wikipedia (+Gnome for CS)
2 In domain	Training	Wikipedia and Europarl
	Test	Wikipedia (+Gnome for CS)
3 Out of domain	Training	Wikipedia and Europarl
	Test	News

Impact: Results on Wikipedia (in domain)

	CS	Sc	Sp	Un
c3g	38.81	40.53	46.94	43.68
cog	57.32	56.17	57.60	58.14
mono$_{en}$	54.27	52.96	55.74	55.17
$\bar{\mathrm{S}}$·len	56.14	57.40	58.39	58.80
union	64.65	62.95	62.65	64.47
Europarl	27.99	34.00	30.02	30.63
EP+c3g	46.07	48.29	50.40	49.34
EP+cog	58.39	57.70	59.05	58.98
EP+mono$_{en}$	54.44	53.93	56.05	55.88
EP+$\bar{S}$·len	56.05	57.53	59.78	58.72
EP+union	66.22	64.24	64.39	65.67

Impact: Results on Gnome (in domain)

	CS	Un
c3g	11.08	9.56
cog	18.48	17.66
mono$_{en}$	19.48	20.58
$\bar{S}$·len	20.71	20.56
union	22.41	20.63
EP	18.15	
EP+c3g	19.78	19.49
EP+cog	21.09	20.14
EP+mono$_{en}$	21.27	20.66
EP+$\bar{S}$·len	21.58	20.65
EP+union	22.37	21.43

Impact: Translation Instances

Better reordering

Source All internet packets have a source IP address and a destination IP address.

EP Todos los paquetes de internet tienen un origen dirección IP y destino dirección IP.

EP+union-CS Todos los paquetes de internet tienen una dirección IP de origen y una dirección IP de destino.

Reference Todos los paquetes de internet tienen una dirección IP de origen y una dirección IP de destino.

Impact: Translation Instances

Awareness of terms (possible overfitting?)

Source Attack of the Killer Tomatoes is a 2D platform video game developed by Imagineering and released in 1991 for the NES.

EP el ataque de los tomates es un asesino 2D plataforma vídeo-juego desarrollados por Imagineering y liberados en 1991 por la NES.

union-CS Attack of the Killer Tomatoes es un videojuego de plataformas desarrollado por Imagineering y lanzado en 1991 para la Nintendo Entertainment System.

Reference Attack of the Killer Tomatoes es un videojuego de plataformas en 2D desarrollado por Imagineering y lanzado en 1991 para el NES.

Impact: Translation Instances

Better vocabulary

Source Fractal compression is a lossy compression method for digital images, based on fractals.

EP Fractal compresión es un método para lossy compresión digital imágenes , basada en fractals.

EP+union-CS La compresión fractal es un método de compresión con pérdida para imágenes digitales, basado en fractales.

Reference La compresión fractal es un método de compresión con pérdida para imágenes digitales, basado en fractales.

Impact: Results on News (out of domain)

	CS	Sc	Sp	Un
union	16.74	22.28	15.82	22.16
Europarl	27.02			
EP+c3g	26.06	26.35	26.81	27.07
EP+cog	26.61	27.33	26.71	27.08
EP+mono$_{en}$	27.18	26.80	26.96	27.44
EP+$\bar{S}$·len	27.59	26.80	27.58	27.22
EP+union	26.76	27.52	27.35	26.72

Outline

Background

Domain-Specific Comparable Corpora Extraction

Parallel Fragments Extraction

Impact on an MT System

Final Remarks

Final Remarks

- A simple model to extract domain-specific comparable corpora from Wikipedia
- The domain-specific corpora showed to be useful to feed SMT systems, but other tasks are possible

Final Remarks

- A simple model to extract domain-specific comparable corpora from Wikipedia
- The domain-specific corpora showed to be useful to feed SMT systems, but other tasks are possible
- We are currently comparing our model against an IR-based system [Plamada and Volk, 2012]
- The platform currently operates in more language pairs, including French, Catalan, German, and Arabic; but it can operate in any language and domain

Final Remarks

- A simple model to extract domain-specific comparable corpora from Wikipedia
- The domain-specific corpora showed to be useful to feed SMT systems, but other tasks are possible
- We are currently comparing our model against an IR-based system [Plamada and Volk, 2012]
- The platform currently operates in more language pairs, including French, Catalan, German, and Arabic; but it can operate in any language and domain
- The prototype is coded in Java (and depends on JWPL). We plan to release it in short!

Thank you!

Alberto Barrón-Cedeño, Cristina España-Bonet,
Josu Boldoba, Lluís Màrquez
QCRI & UPC
albarron@qcri.org.qa cristinae@cs.upc.edu

References I

McNamee, P. and Mayfield, J. (2004).
Character N-Gram Tokenization for European Language Text Retrieval.
Information Retrieval, 7(1-2):73–97.

Plamada, M. and Volk, M. (2012).
Towards a wikipedia-extracted alpine corpus.
In The Fifth Workshop on Building and Using Comparable Corpora.

Pouliquen, B., Steinberger, R., and Ignat, C. (2003).
Automatic Identification of Document Translations in Large Multilingual Document Collections.
In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003), pages 401–408, Borovets, Bulgaria.

Simard, M., Foster, G. F., and Isabelle, P. (1992).
Using Cognates to Align Sentences in Bilingual Corpora.
In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation.