Domain-specific terminology extraction for Machine Translation

Mihael Arcan

Outline

- Phd topic
- Introduction
- Resources
- Tools
- Multi Word Extraction (MWE) extraction
- Projection of MWE
- Evaluation
- Future Work
- Conclusion

Phd topic

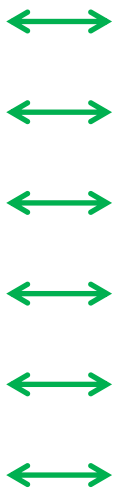
- Translating of domain-specific terms
- Issue: Out-Of-Vocabulary, specific structure of terms, OOV brings along ambiguity
- Possible solutions:
  - Generating new parallel resources
  - Using specific vocabulary and embed it correctly into the SMT system

Introduction

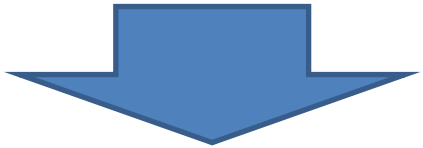
- *"Domain-specific terminology extraction for MT"*
  - How to extract domain-specific terminology?
    - what is a domain-specific term?
    - what is a domain?
  - How to embed the extracted terminology into SMT?
    - how much data we need to improve MT
    - using existing resources
    - adding terminology to general resources

Introduction

Permission is hereby granted, **free of charge**, to any person obtaining a copy of the **Unicode data files** to deal in them without restriction, including without limitation the rights to **use**, **copy**, modify, merge, publish, distribute, and/or sell copies.



A chiunque ottenga una **copia** dei **file di dati Unicode** è concesso il permesso, senza oneri, di distribuirli senza restrizioni, ivi incluso senza limitazione dei diritti di uso, copia, modifica, unione, **pubblicazione**, **distribuzione** e/o vendita di copie.



KDE is available **free of charge**, but costs are incurred and assets formed in its creation. Thus, the KDE community ...

Resources

- JRC – Acquis
  - http://ipsc.jrc.ec.europa.eu/index.php?id=198
- Gnome
  - https://l10n.gnome.org
  - 66 documents (10 dev / 56 test set)
    - 166 sentences for extracting keyphrases
- KDE4
  - http://opus.lingfil.uu.se/
  - Top 1000 sentences with most MWE overlap

kx Toolkit[1] for Multilingual MWE extraction

- kx Toolkit:
  - n-gram based extraction
  - POS annotation
  - tf-idf information
- (Parallel) corpus for the source and target language
  - In-domain data (gnome in English and Italian)
  - Out-domain data (JRC Acquis in English and Italian)

(1) http://dl.acm.org/citation.cfm?id=1859700

kx Toolkit[1] for Multilingual MWE extraction

- English keyphrase extraction:
  - accelerates, activation requests, active applications, active, antialiasing, application manager, available zoom levels, desktop, detailed metadata list, development packages, ...

Italian keyphrase extraction:

- accelera, richiesta di attivazione, applicazioni attive, attivo, antialiasing, gestore applicazioni, livelli di ingrandimento disponibili, elenco dettagliato dei metadati, pacchetti di sviluppo, ...

kx keyphrase extraction - Evaluation

- Annotation of 10 gnome files

  A {1 string} specifying how parts of overlong {2 file names} should be replaced by {3 ellipses}, depending on the {4 zoom level}.

  A {1 string} specifying how parts of overlong {2 file names} should be replaced by ellipses, depending on the {3 zoom level}.

- F-score
  - 0.38 (p0.35/r0.42) for English and 0.40 (p0.46/r0.35) for Italian

Projection Options

- Word alignment
  - Giza++
- PB-SMT alignment
  - Moses decoder
  - SRILM Toolkit

Word Alignment Projection Options

- Option 1: Word Alignment
- Option 2: Option 1 + Contiguous Span
- Option 3: Option 1 + Sentence Lookup
- Option 4: Option 1 + kx lookup

Word Alignment Projection Options

- Option 1: Word Alignment

titlebar-font-size option ||| opzione titlebar-font-size

size used ||| dimensione usata

email systems ||| sistemi di email

executable text files ||| file testo eseguibili

*order **of** subpixel elements ||| ordine elementi subpixel

Word Alignment Projection Options

- Option 2: Option 1 + Contiguous Span

order of subpixel elements ||| ordine degli elementi subpixel

protection of personal data ||| protezione dati personali

Word Alignment Projection Options

- Option 3: Option 1 + Sentence Lookup

error ||| errore

embed ||| incorporata

straight-line method of depreciation ||| metodo di ammortamento lineare

given number of lines ||| numero di righe indicato

Word Alignment Projection Options

- Option 4: Option 1 + kx lookup

specified zoom level ||| livello di ingrandimento specificato

activation requests ||| richiesta di attivazione

webmmux ||| webmmux

shortcuts ||| *predefinite

(*scorciatoia*, but we get *default*)

PB-SMT Projection Options

- Option 1: PB-SMT best translation
- Option 2: Option 1 + Sentence Lookup
- Option 3: Option 1 + kx Lookup
- Option 4: n-best PB-SMT + Sentence Lookup
- Option 5: n-best PB-SMT + kx Lookup

(n=10)

Bilingual Projections

	In-domain	Excluded by Out
Word Alignment (WA)	356	190
WA + Contiguous Span	352	190
WA + Sentence lookup	284	186
WA + kx lookup	115	29
PB-SMT	1015	126
PB-SMT + Sent. lookup	549	42
PB-SMT + kx lookup	143	20
Nbest PB-SMT + Sent.	549	80
Nbest PB-SMT + kx lookup.	157	46

Bilingual Projections

	In-domain	Precision
Word Alignment (WA)	356	0.75
WA + Contiguous Span	352	0.65
WA + Sentence lookup	284	0.8
WA + kx lookup	115	0.85
PB-SMT	1015	0.35
PB-SMT + Sent. lookup	549	0.7
PB-SMT + kx lookup	143	0.9
Nbest PB-SMT + Sent.	549	0.7
Nbest PB-SMT + kx lookup.	157	0.9

Embedding Terminology into SMT

- JRC Acquis
- Acquis + Gnome parallel data
- Acquis + kx keyphrases

- Xml markup
- Xml markup with CacheBased LM
- JRC Acquis Cache Based TM + Cache Based LM

Translation Models- Evaluation

	BLEU
JRC Acquis	14.62
XML Markup WA with kx lookup	15.23
XML Markup WA with sentence	13.39
XML Markup WA with contiguous span	11.99
XML Markup n-best PB-SMT w. kx lookup	14.94
XML Markup n-best PB-SMT w. sentence lookup	14.69
XML Markup n-best PB-SMT without proving	14.31

Translation Models- Evaluation

	BLEU
JRC Acquis	14.62
CBXL Markup WA with kx lookup	15.26
CBXL Markup WA with sentence	14.29
CBXL Markup WA with contiguous span	14.53
CBXL Markup n-best PB-SMT w. kx lookup	14.93
CBXL Markup n-best PB-SMT w. sentence lookup	14.61
CBXL Markup n-best PB-SMT without proving	14.14

Future Work

- Analyzing the results
- Applying the experiment to

  English->French and English->German
- Extended bilingual evaluation
- Experiments on other domains
  - Additional experiments on close domains
- Using comparable data for keyphrase extraction

  (e.g. Wikipedia)

Conclusion

- Extraction of domain-specific MWE
- Bilingual Projection of MWE
- Novel results of embedding the terminological knowledge into SMT

Thank you