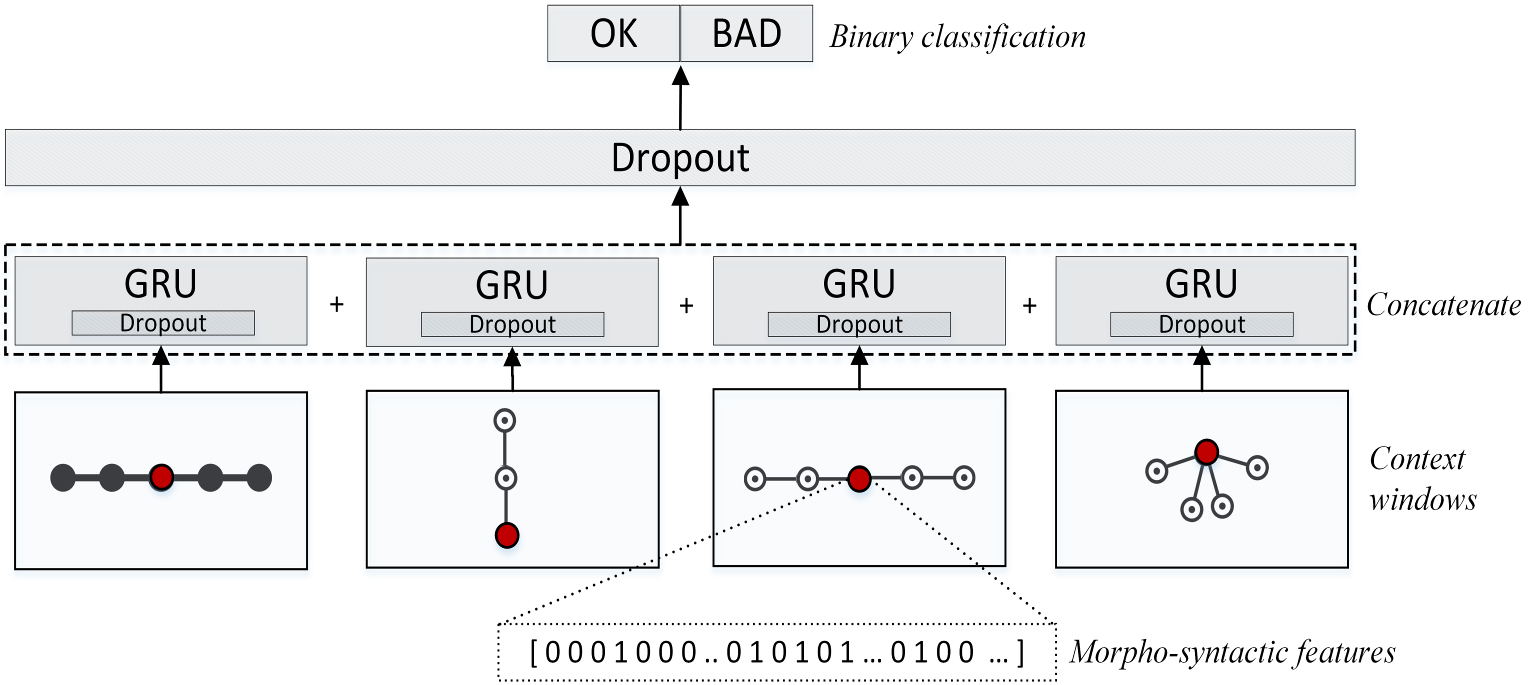
A Neural Network Architecture for Detecting Grammatical Errors in SMT



A Neural Network Architecture for Detecting Grammatical Errors in SMT

- Morpho-Syntactic features outperform word embeddings on this task
- Syntactic n-grams improve the performance
- This method can successfully be applied
  - across languages
  - to detect post-editing effort

Evaluating the Usability of a Controlled Language Authoring Assistant

Rei Miyata (Nagoya U.), Anthony Hartley (Rikkyo U.), Kyo Kageura (U. of Tokyo), Cécile Paris (CSIRO)

Improved machine translatability when a controlled language (CL) is employed

→ Two sets of Japanese CL rules for RBMT and SMT (Total: 36 rules)

Practical problem: Difficulty in manually applying a number of CL rules

Variant (incorrect) term

Rule 18: particle *Ga* (が) for object

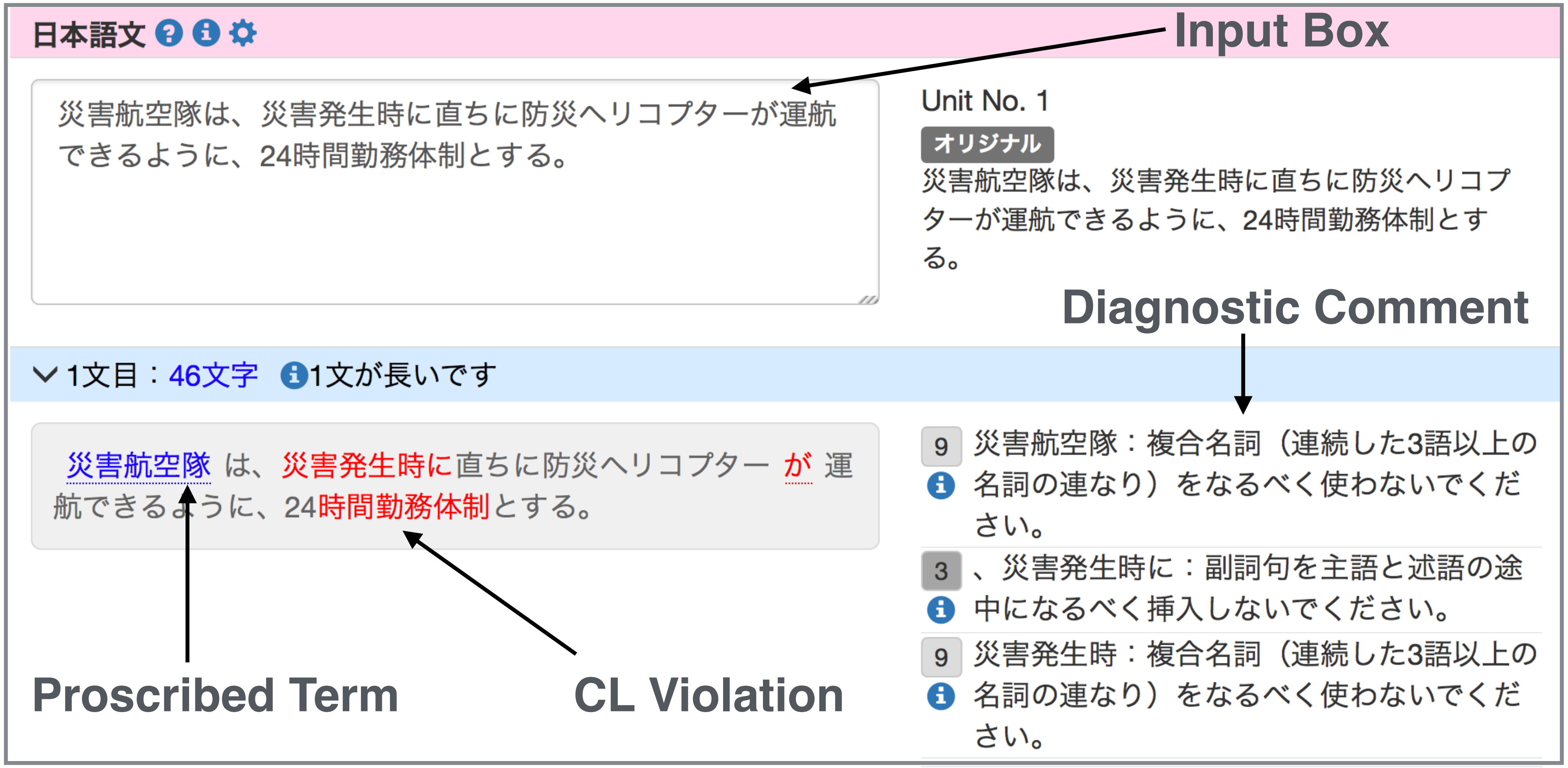
災害航空隊は、災害発生時に直ちに防災ヘリコプターが運航できるように、24時間勤務体制とする。

Rule 28: compound word

Rule 20: inserted adverbial clause

[Reference] The Disaster Prevention Fleet has **a 24-hour duty system** so that they can operate their emergency helicopters promptly if a disaster occurs.

Solution: CL authoring assistant for non-professional writers



Web-based

Real-time

Interactive

Ja-En MT

30 CL rules

Municipal domain

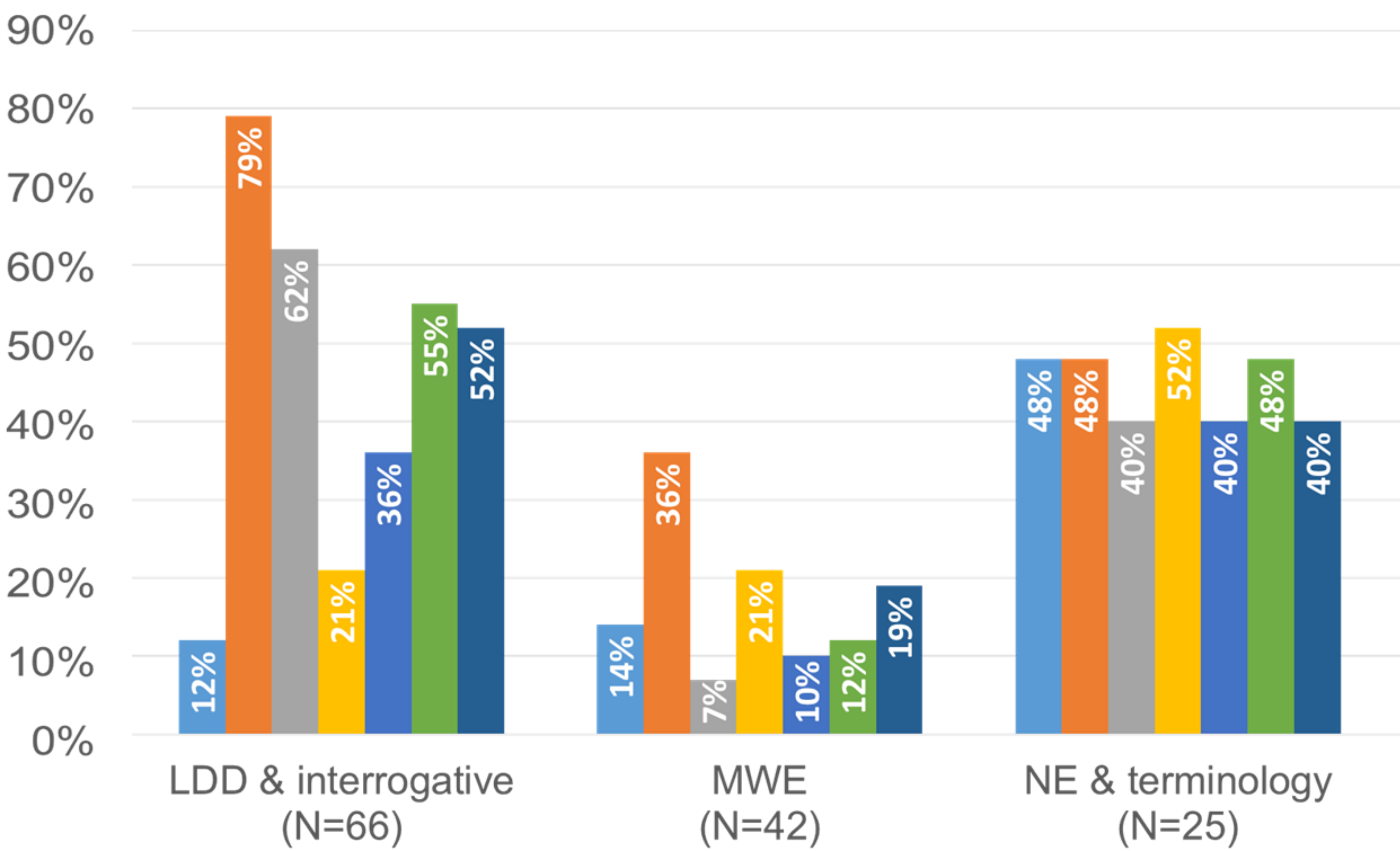
How usable our system is?

Effectiveness	Does the system help reduce CL violations and improve MT quality?	✓
Efficiency	Does the system help reduce time spent on controlled writing?	✓
Satisfaction	Is the system easy for non-professional writers to use and favourably accepted?	✓

Linguistic-Driven Evaluation of MT Output

- Test suites have been a familiar tool in NLP in areas such as grammar development
- Why not use test suites in MT development?
- **Our approach**
  - Manual creation of comprehensive test suite (~ 5,000 test items per language direction)
  - Testing of 7 different MT systems on a subset of the test suite: 1 RBMT, 2 PBMT, 4 NMT

Sneak Peek of Results



Pre-Reordering for Neural Machine Translation: Helpful or Harmful?

- **Consensus on NMT & SMT**
  - NMT produces more fluent translations than SMT
  - NMT produces more changes in the word order of a sentence
  - Pre-reordering is helpful to SMT
- **A Straightforward Question**
  - Is pre-reordering also helpful to NMT?
- **Intuitional Contradiction:**
  - Pre-reordering is necessary: it can facilitate the attention mechanism to learn a diagonal alignment
  - Pre-reordering is redundant: the attention mechanism is capable of globally learning the word alignment
- **What is the truth?!**

Pre-Reordering for Neural Machine Translation: Helpful or Harmful?

- **Findings from NMT pre-reordering exepriment**
  - Pre-reordering deteriorates translation performance of NMT systems
  - Pre-reordered NMT is better than non-reordered SMT, but worse than pre-reordered SMT
- **How does the pre-reordering contribute to NMT?**
  - Pre-reordering features as input factors for NMT
- **Does it work?**
  - Yes, it works!
  - Please come to our poster for more!

- We **need** to post-edit MT output for dissemination purposes and this is expensive
- So why don't we directly **optimize MT systems to improve their usefulness in post-editing**?
- It makes sense to use extensive metrics to evaluate MT: how many euros, hours, edits...?
- We study a collection of metrics and evaluate their performance in predicting **post-editing effort**
- Can good-old *BLEU* still be a good metric for this task?

find it out at our poster!

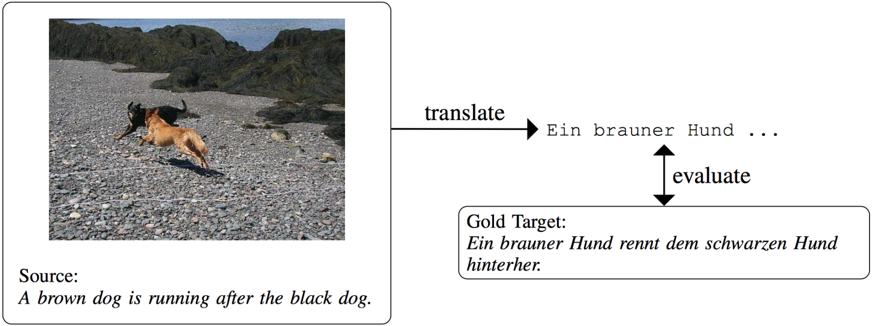
Towards Optimizing MT for Post-Editing Effort: Can BLEU Still Be Useful?

•••

Mikel L. Forcada,[1] Felipe Sánchez-Martínez,[1] Miquel Esplà-Gomis,[1] Lucia Specia[2]

[1]Universitat d'Alacant — [2]Sheffield University

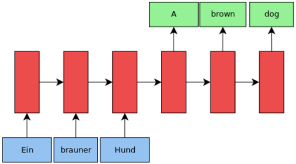
Unraveling the Contribution of Image Captioning and Neural Machine Translation for Multimodal Machine Translation



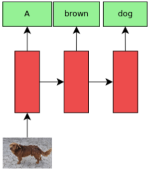
Given an image description in a source language and its corresponding image, translate it into a target language

Our Contribution

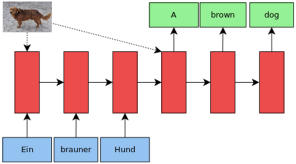
- We isolate two distinct but related components of Multimodal Machine Translation and analyse their individual contributions



(a) Neural MT



(b) IC



(c) Multimodal NMT

- We propose a method to combine the outputs of both components to improve translation quality

Reference	a dog treads through a shallow area of water located on a rocky mountainside.
Baseline	a dog walks through a body of water, with a body of water in it.
AVERAGE	a dog walks through a body of water, looking at a rocky mountain.

Comparing Language Related Issues for NMT and PBMT between German and English

– Maja Popović –

- German is a complex language for (phrase-based) machine translation
- NMT yields large improvements of automatic evaluation scores in comparison to PBMT
  - especially for English→German

- related work on more detailed (automatic) evaluation and error analysis:
  - NMT mainly improves fluency, especially reordering
  - adequacy not clear
  - long sentences (>40 words) not clear

This work (manual analysis):

- what particular language related aspects (issues) are improved by NMT?
  - $\rightarrow$ definitely several aspects of fluency (grammar)
- are there some prominent issues for NMT itself?
  - $\rightarrow$ yes, there are
    only adequacy? not sure
- are there complementary issues?
  i.e. is combination/hybridisation worth investigating?
  - $\rightarrow$ yes

HOW TO: Make a fully-functioning postedition-quality MT system from scratch using only

- Sophisticated neural wetware
- Billions of neurons
- Zero hidden layers

Find out how *this group* did it with one simple trick!

Rule-based machine translation for the Italian–Sardinian language pair

Francis M. Tyers,[1,2] Hèctor Alòs i Font,[3]
Gianfranco Fronteddu,[4] and Adrià Martín-Mor.[5]

[1] UiT Norgga árktalaš universitehta;
[2] Tartu ülikool;
[3] Universitat de Barcelona;
[4] Università degli studi di Cagliari;
[5] Universitat Autònoma de Barcelona

Continuous Learning from Human Post-edits for Neural Machine Translation

M.Turchi, M. Negri, M.A. Farajian and M. Federico

Expectation

Reality

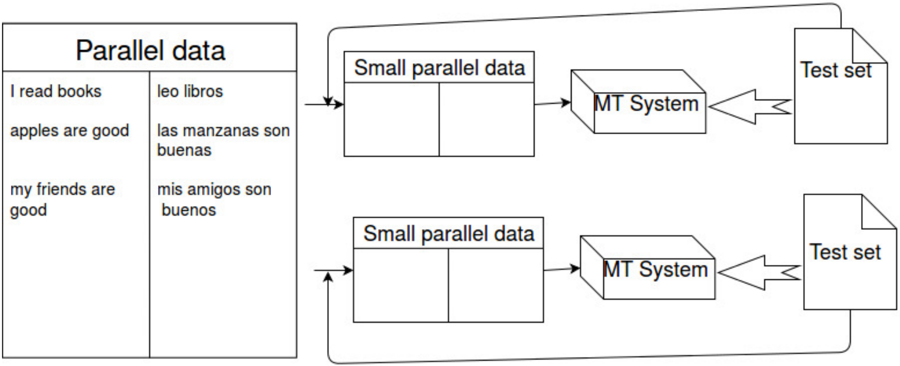
Continuous Learning from Human Post-edits for Neural Machine Translation

M.Turchi, M. Negri, M.A. Farajian and M. Federico

Feedback can help… …but

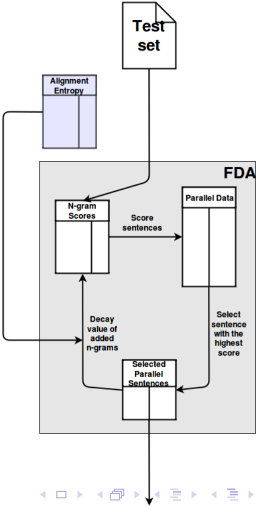
Applying N-gram Alignment Entropy to Improve Feature Decay Algorithms

Data selection task



Applying N-gram Alignment Entropy to Improve Feature Decay Algorithms

- Use of FDA.
- Use of entropies to make parameters of FDA dynamic.



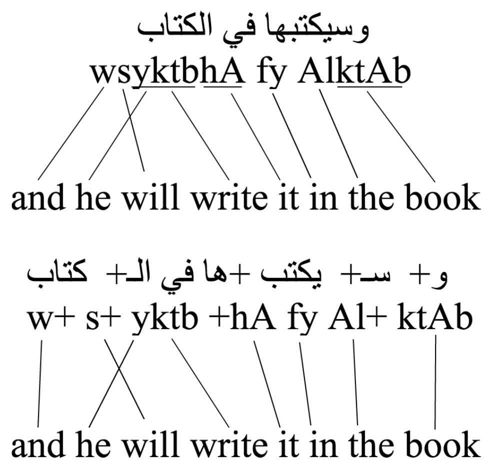
Optimizing Tokenization Choice for Machine Translation across Multiple Target Languages

Nasser Zalmout and **Nizar Habash**

New York University Abu Dhabi, UAE

{nasser.zalmout,nizar.habash}@nyu.edu

Tokenization is good for machine translation…



Tokenization Scheme		Example	
D0	no tokenization	wsyktbhA	llTAlb
D1	split CONJ	w+ syktbhA	llTAlb
D2	split CONJ and PART	w+ s+ yktbhA	l+ AlTAlb
ATB	Arabic Treebank	w+ s+ yktb +hA	l+ AlTAlb
D3	split all clitics	w+ s+ yktb +hA	l+ Al+ TAlb

Tokenization schemes **work as blueprint for the tokenization process, controlling the intended level of verbosity**

The tokenization scheme choice for Arabic, is typically *fixed* for the whole source text, and *does not vary with the target language*

This raises many questions:

- Would the best source language tokenization choice vary for different target languages?
- Would combining the various tokenization options in the training phase enhance the SMT performance?
- Would considering different tokenization options at decoding time improve SMT performance?

We use Arabic as source language, with five target languages of varying morphological complexity: English, French, Russian, Spanish, and Chinese

Sounds interesting? Come to our poster!

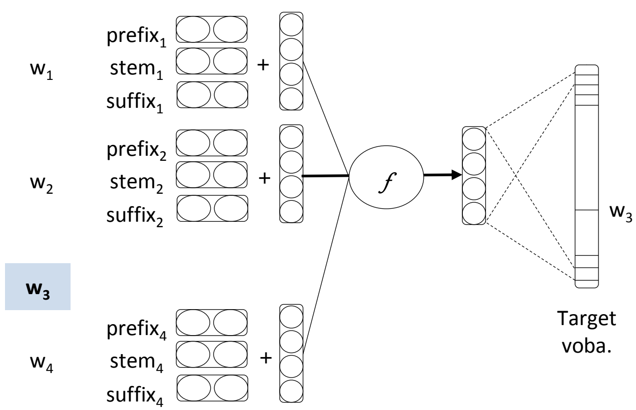
Peyman Passban, Qun Liu and Andy Way

Introduction

Farsi (Persian) is a low resource and morphologically rich language and it is quite challenging to achieve acceptable translations for this language. Our goal is to boost existing SMT models for Farsi via auxiliary morphological information provided by neural networks (NNs). To this end we propose two solutions:

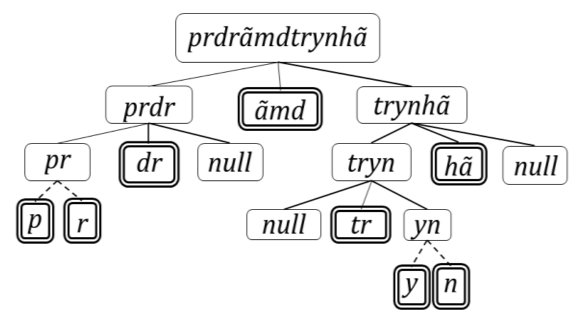
- We introduce an additional morphological factor for the factored SMT model.
- We substitute the existing n-gram-based language model with a subword-aware neural language model.

Neural Model for training
Morphology-aware Embeddings



$$\varepsilon(w_i) = \varepsilon(pre_i) + \varepsilon(stm_i) + \varepsilon(sfx_i) + \varepsilon(w_i)$$

Segmentation Model for
Decomposing Complex Words



Direction	Baseline	Extend_3	Extend_4^w	Extend_4^m
En→De	21.11	21.42	21.57	**21.70**
De→En	29.50	29.58	29.71	**29.78**
En→Fa	21.03	22.14	22.27	**22.61**
Fa→En	29.21	30.53	30.67	**30.91**

+1.58

Direction	Baseline	n-gram^w	n-gram^m
En→De	21.11	21.53	**21.88**
De→En	29.50	29.87	**30.43**
En→Fa	21.03	21.86	**22.36**
Fa→En	29.21	29.91	**31.05**

+1.33

Model	German (De)	Farsi (Fa)
Botha and Blunsom (2014)	296	-
Kim et al. (2016)	239	128
Proposed Model	**225**	**110**

Neural Networks Classifier for Data Selection in Statistical Machine Translation

Á. Peris*, M. Chinea-Rios*, F. Casacuberta*

*PRHLT Research Center{*lvapeab,machirio, fcn*}*@prhlt.upv.es*

May 26, 2017

Main contributions of this work

- We tackle the DS problem for SMT as a classification task employing CNNs and bidirectional long short-term memory (BLSTM) networks.
- Introduce two architecture of the proposed classifiers (Monolingual and Bilingual).
- Present a semi-supervised algorithm for training our classifiers.
- The results show that our method outperforms a state-of-the-art DS technique in terms of translation quality and selection sizes.
- We show that both CNNs and BLSTM networks provide a similar performance for the task at hand.

Historical Documents Modernization

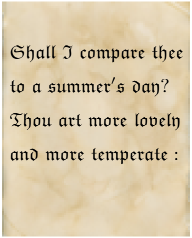
Miguel Domingo, Mara Chinea-Rios, Francisco Casacuberta

midobal@prhlt.upv.es, machirio@prhlt.upv.es, fcn@prhlt.upv.es

Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València

EAMT 2017

Prague, May 31, 2017


Shall I compare thee
to a summer's day?
Thou art more lovely
and more temperate :

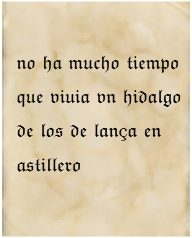
Original document

Transcription

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:

Shall I compare you to a summer day?
You're lovelier and milder.

Modern version


no ha mucho tiempo
que viuia vn hidalgo
de los de lança en
astillero

Original document

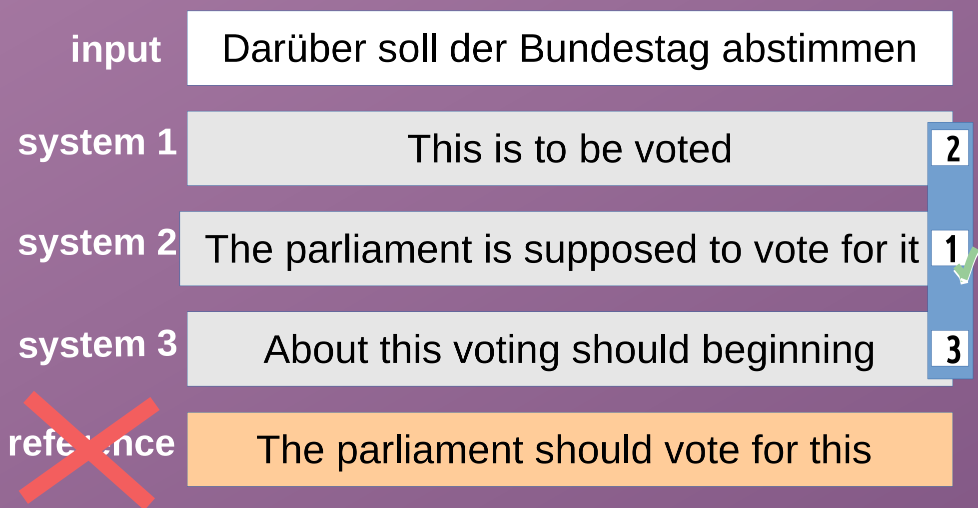
Transcription

no ha mucho tiempo que viuia vn hidalgo
de los de lança en astillero

no ha mucho tiempo que vivía un hidalgo
de los de lanza en astillero

Version with updated spelling

Comparative Quality Estimation



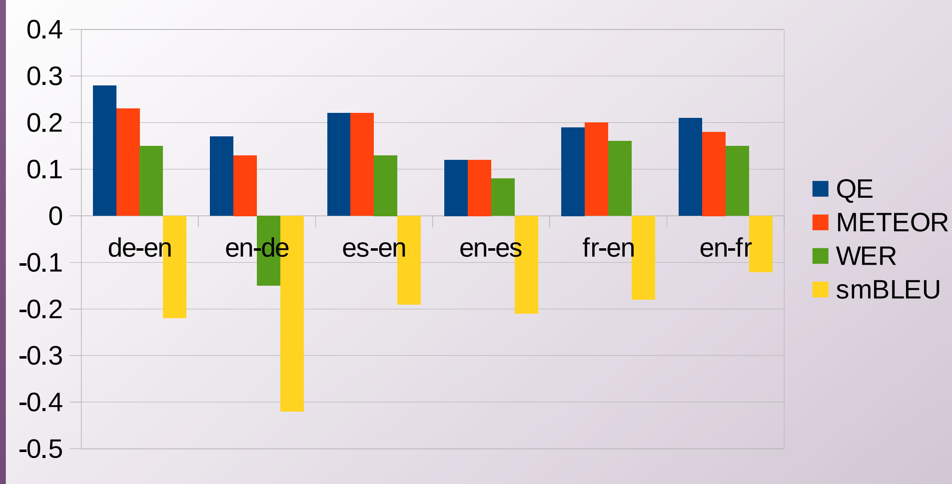
Machine learning to **compare** alternative translations

- focus on one sentence at a time
- one source sentence with many translations
- don’t use reference
- rank translations (best to worse)

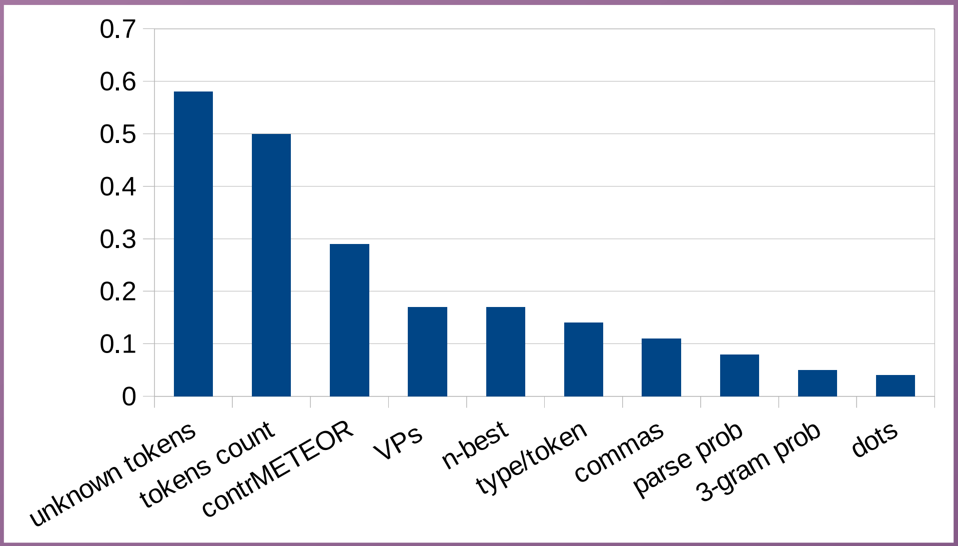
new learner: Gradient Boosting

features: introduce adequacy features add more fluency features

- Applied on WMT output from 7 years, 6 language directions
- Beats automatic metrics. → ML better than references



Comparative Quality Estimation



Feature conclusions

– **Target fluency** (grammatical) features are important

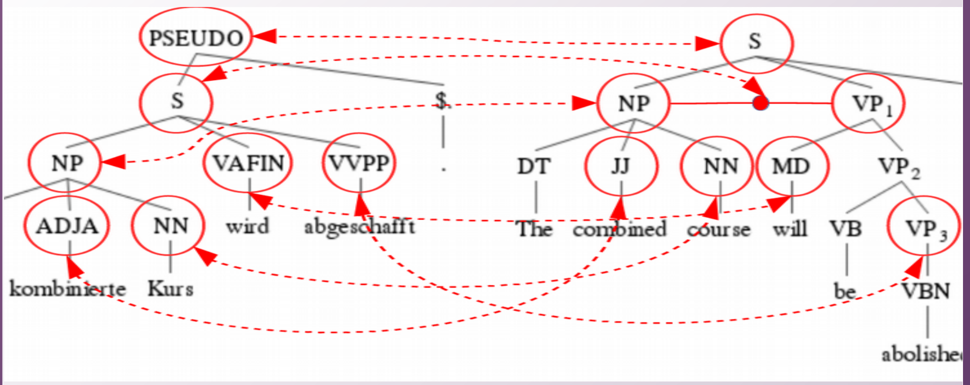
– Few **adequacy** features are useful

– Source complexity features are useless

Language specific observations

– en-de: position of the VPs and PPs

– de-en: count of CFG rules with noun determiners, gerunds, PPs with “in”

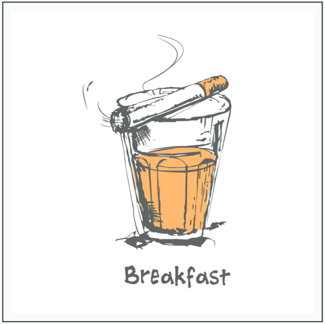


This work has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 645452

Problem: going from the top to the bottom to translate important conversations.

tujhyasathi	gold	ani	cutting	aanto
tujhyāsāṭhī	golḍ	āṇi	kaṭiṃg	āṇato
तुझ्यासाठी	गोल्ड	आणि	कटिंग	आणतो

“I’ll get you a cigarette and tea”



Finite-state back-transliteration for Marathi

Vinit Ravishankar

University of Malta

Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English

Duygu Ataman, Matteo Negri, Marco Turchi, Marcello Federico

PROBLEM

- Sub-word segmentation approaches in NMT can disrupt the semantic and syntactic structure of agglutinative languages like Turkish

Source	Segmentation	NMT Output	Reference
kanunda	kan@@ unda	in **your blood**	in **the law**
sigortalılar	sigor@@ talı@@ lar	the insure**rs**	the insure**d ones**

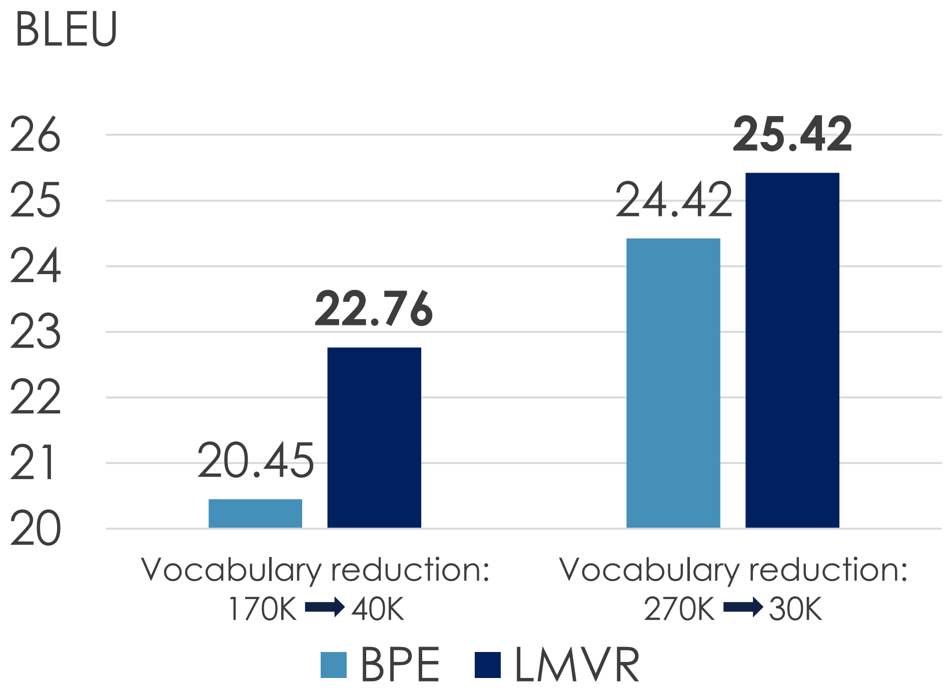
Translation examples obtained when Byte-Pair Encoding is applied on Turkish words

Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English

Duygu Ataman, Matteo Negri, Marco Turchi, Marcello Federico

SOLUTION

- Linguistically Motivated Vocabulary Reduction (LMVR)
  - Considers morphological properties of the sub-word units
  - Controls vocabulary size during segmentation
  - Unsupervised algorithm which can be used in other languages



Questing for Quality Estimation
A User Study

Carla Parra Escartín[1], Hanna Béchara[2], Constantin Orăsan[2]

[1] ADAPT Centre, SALIS, Dublin City University, Ireland
[2] RGCL, University of Wolverhampton, UK

Does MTQE **really** help translators?

- 4 translators EN→ES
- 1 MTPE task, 300 sentences and 4 conditions:

Translate	Post-Edit
MTQE says… Post-Edit!	MTQE says… Translate!

If you want to see what we found out, come to our poster ;-)

Improving Machine Translation through Linked Data

BabelNet

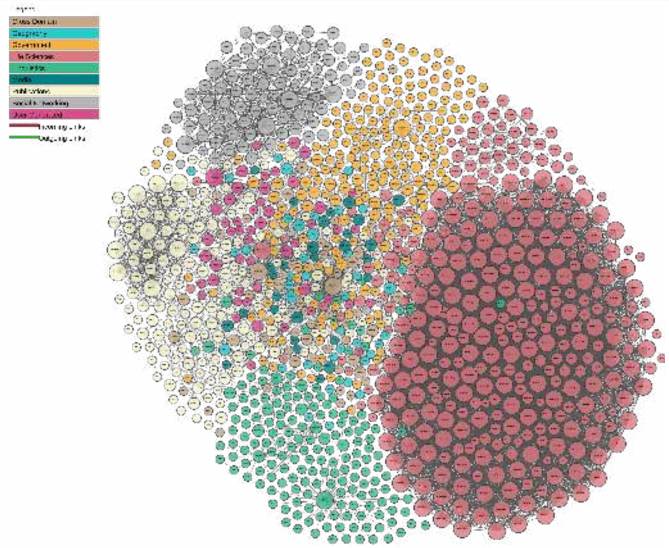
14M entries
271 languages

DBpedia

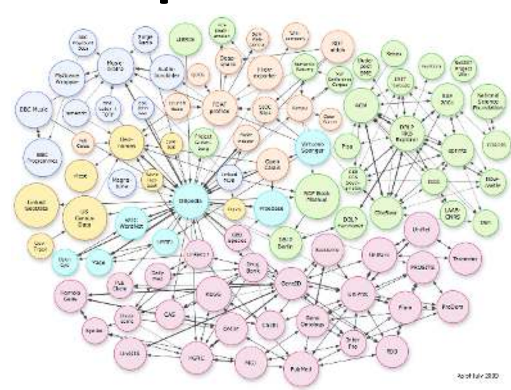
4.58M entries
125 languages

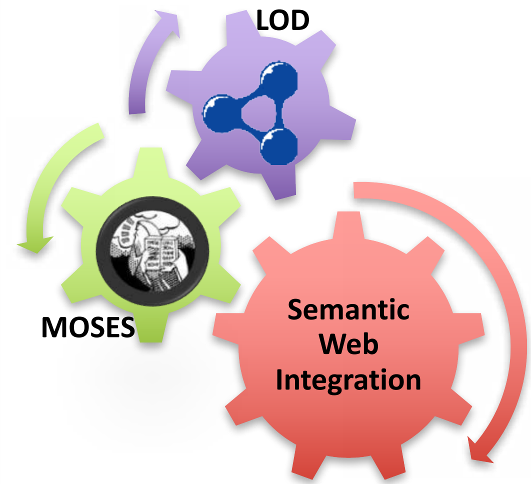
Joint Research Centre JRC

205K entries
20 languages



Linked Open Data Cloud





3 Algorithms:

- Dictionaries
- Pre-Decoding
- Post-Processing

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

DKT

[Srivastava et al., 2017 - EAMT]

WK·POTENZIAL UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative Neue Länder

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Improving Machine Translation through Linked Data

Experimental Set Up

- **Moses** Statistical Machine Translation: English – {German | Spanish}
- 3 Linked Data Resources: **DBpedia** | **BabelNet** | **JRC-Names**

Improving MT Outputs

- Translating **Named Entities** via SPARQL Queries as Decoding Rules
- Translating **Unknown Words** during Post-Editing

Benefits To Community

- Application of freely available online **Multilingual Datasets**
- Making Machine Translation **Semantic Web-aware**

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

DKT

[Srivastava et al., 2017 - EAMT]

WK · POTENZIAL
UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative Neue Länder

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung