University of Zurich[UZH]

Institute of Computational Linguistics

Treatment of Markup in Statistical Machine Translation

Master's Thesis

Mathias Müller

September 13, 2017

Motivation

Researchers often give themselves the luxury of pretending that only pure text matters [...] (Joanis et al., 2013)

- tendency to overlook “applied” problems

One of those applied problems: markup handling

Ich bitte Sie, sich zu einer Schweigeminute zu **erheben**.

One of those applied problems: markup handling

Ich bitte Sie, sich zu einer Schweigeminute zu **erheben**.

Ich bitte Sie, sich zu einer Schweigeminute zu erheben.

One of those applied problems: markup handling

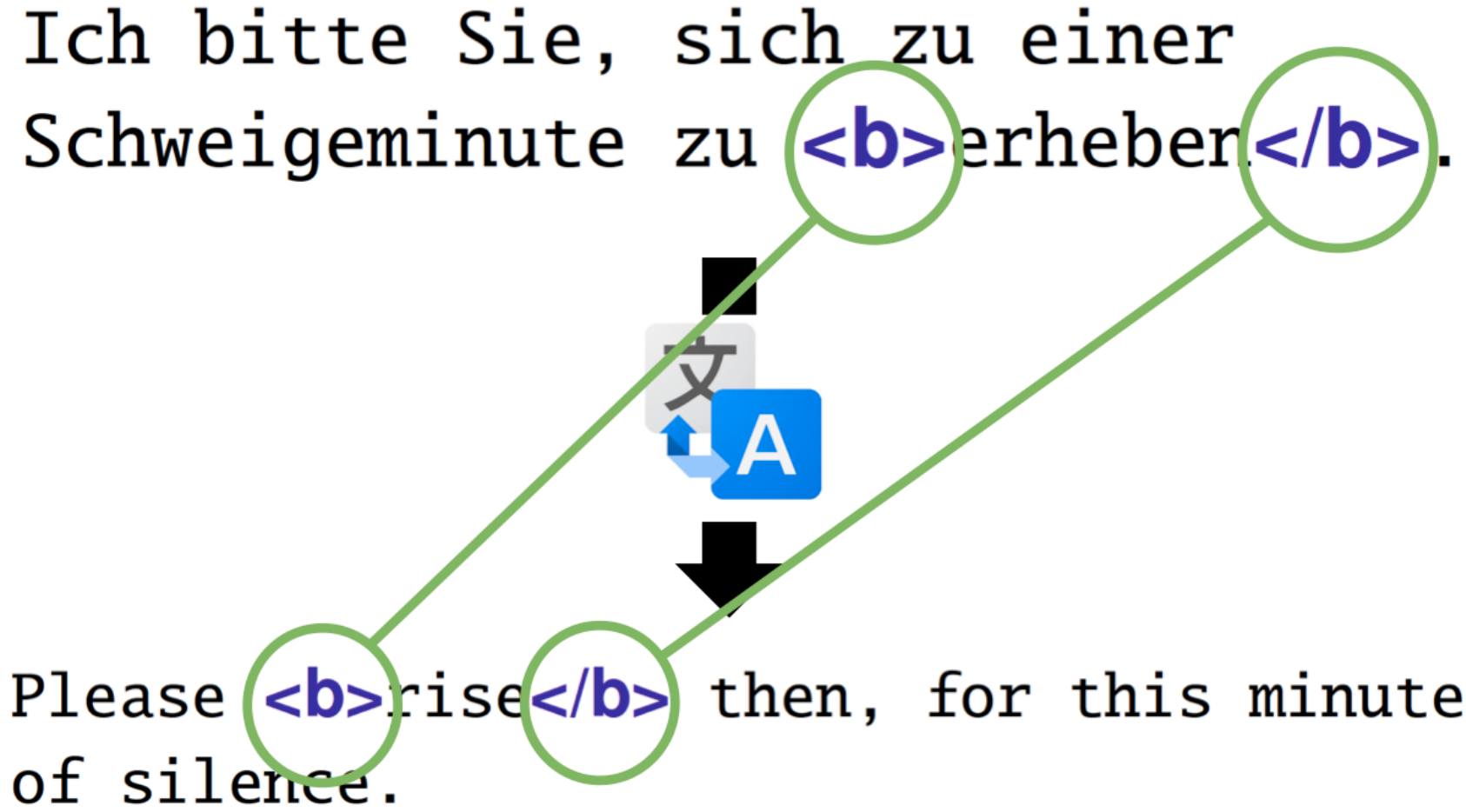
Ich bitte Sie, sich zu einer Schweigeminute zu erheben.

One of those applied problems: markup handling

Ich bitte Sie, sich zu einer Schweigeminute zu erheben.

Please rise then, for this minute of silence.

The essence of markup handling



Why is markup handling important?

(1) Translation industry has a need to translate content with markup

(2) Markup errors harm translator productivity[1]

[1] see e.g. O'Brien (2011); Joanis et al. (2013); Parra Escartín and Arcedillo (2015)

Existing solutions

Complete list of popular machine translation frameworks that have markup handling:

-

Existing solutions

- Is the code that implements markup handling available to the public?
- Are there empirical comparisons of the proposed method?

publication	code	experiments
Du et al. (2010)	✗	✓
Zhechev and van Genabith (2010)	✗	✗
Hudík and Ruopp (2011)	✓	✗
Tezcan and Vandeghinste (2011)	✗	✓
Joanis et al. (2013)	✗	✗

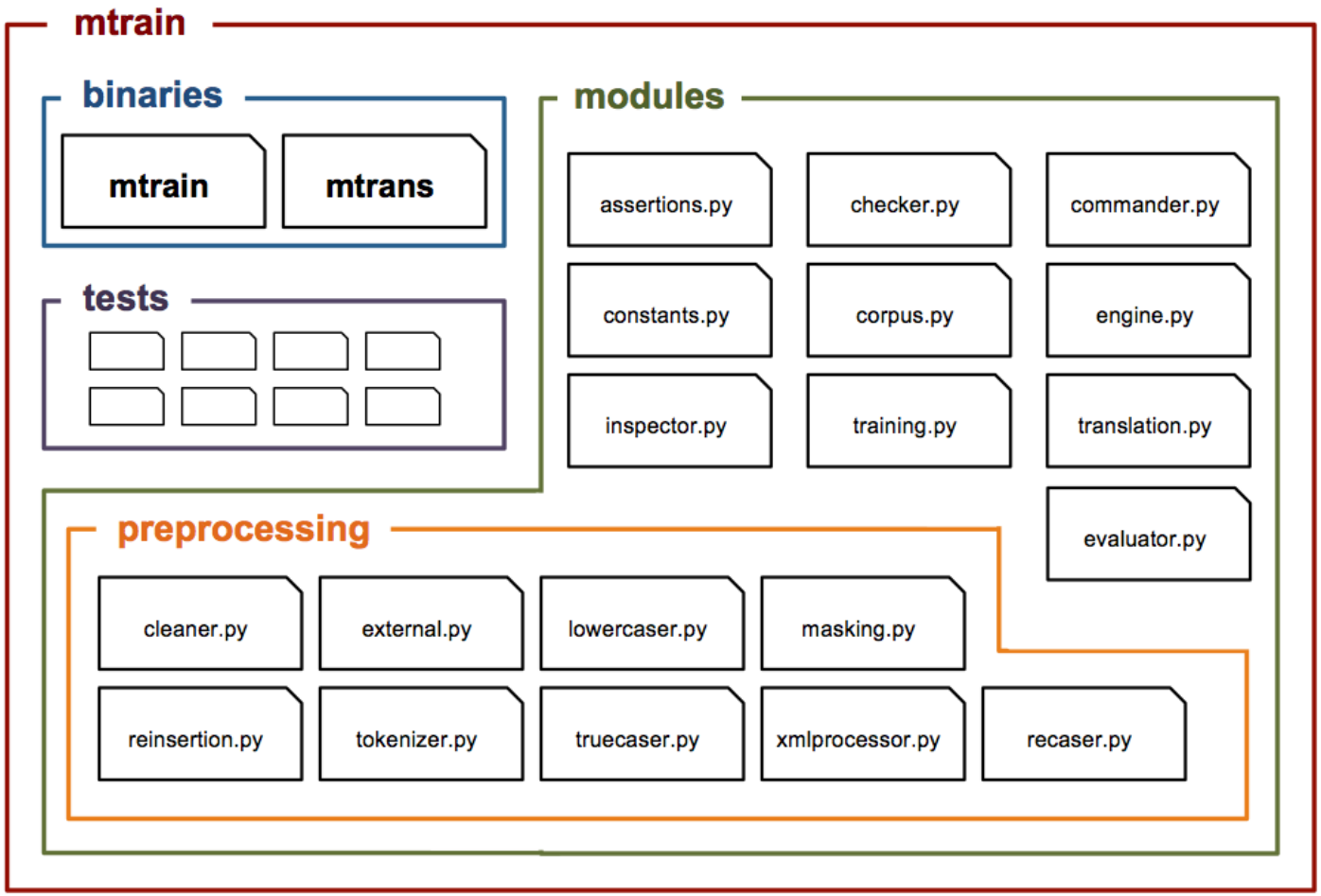
Existing solutions

- Is there pseudo code or a clear explanation of the algorithm?

solution	code	experiments	algorithm
D (2010)	✗	✔	✔
Z (2010)	✗	✗	✔
H (2011)	✔	✗	✔
T (2011)	✗	✔	✗
J (2013)	✗	(✔)	✔

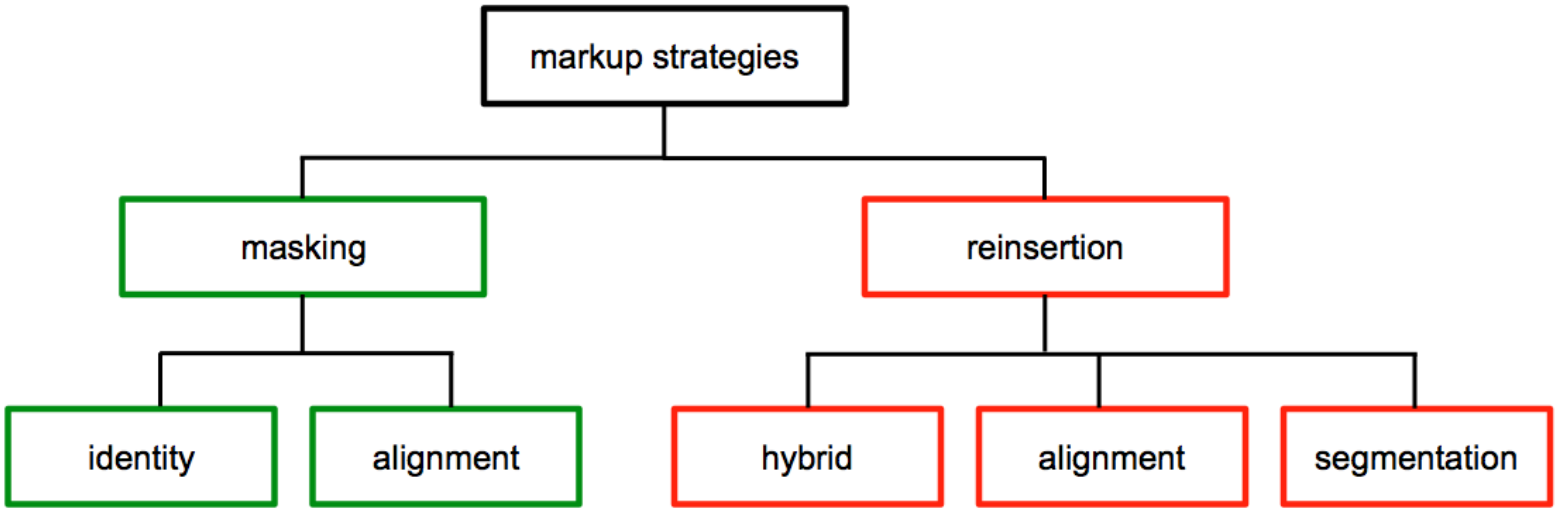
solution	code	experiments	algorithm
thesis work	✔	✔	✔

The `mtrain` framework



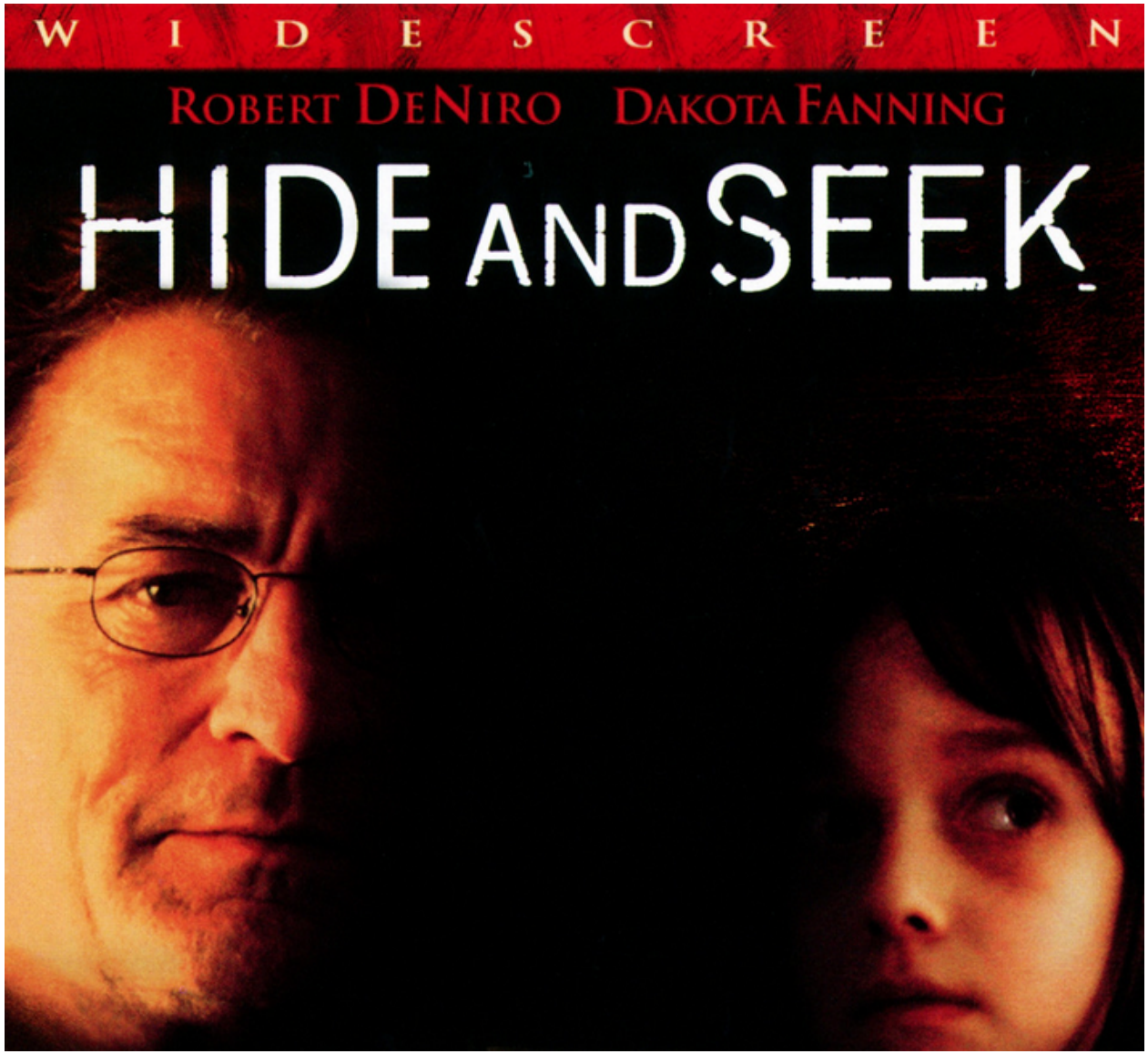
Code: `https://gitlab.cl.uzh.ch/mt/mtrain`

Implemented strategies



- two general classes of strategies, **masking** and **reinsertion**
- how do they solve the problem of markup handling?

How to solve markup handling?



How to solve markup handling?

Make tokenization more sophisticated.

After that, **hide**:

- replace markup with an innocuous string (**masking**)
- remove markup alltogether (**reinsertion**)

... in any case, the original content needs to be **restored** after translation (**seek**)

How Masking Works

List of British jokes updated by rowan@kinson.com

For business inquiries call +4140356 12 25.

<bpt id="1" ctype="bold">{}</bpt >Fios2<ept id="1">{}</ept>

Liste von britischen Witzen aktualisiert von rowan@kinson.com

Kontakt für Firmen: +4140356 12 25.

<bpt id="1" ctype="bold">{}</bpt >Fios2<ept id="1">{}</ept>

How Masking Works

List of British jokes updated by rowan@kinson.com

For business inquiries call +4140356 12 25.

<bpt id="1" ctype="bold">{}</bpt>Fios2<ept id="1">{}</ept>

Liste von britischen Witzen aktualisiert von **EMAIL**

Kontakt für Firmen: **TEL**.

XML {} **XML** Fios2 **XML** {} **XML**

How Masking Works

This is what the final training corpus looks like!

How Reinsertion Works

List of British jokes updated by rowan@kinson.com

For business inquiries call +4140356 12 25.

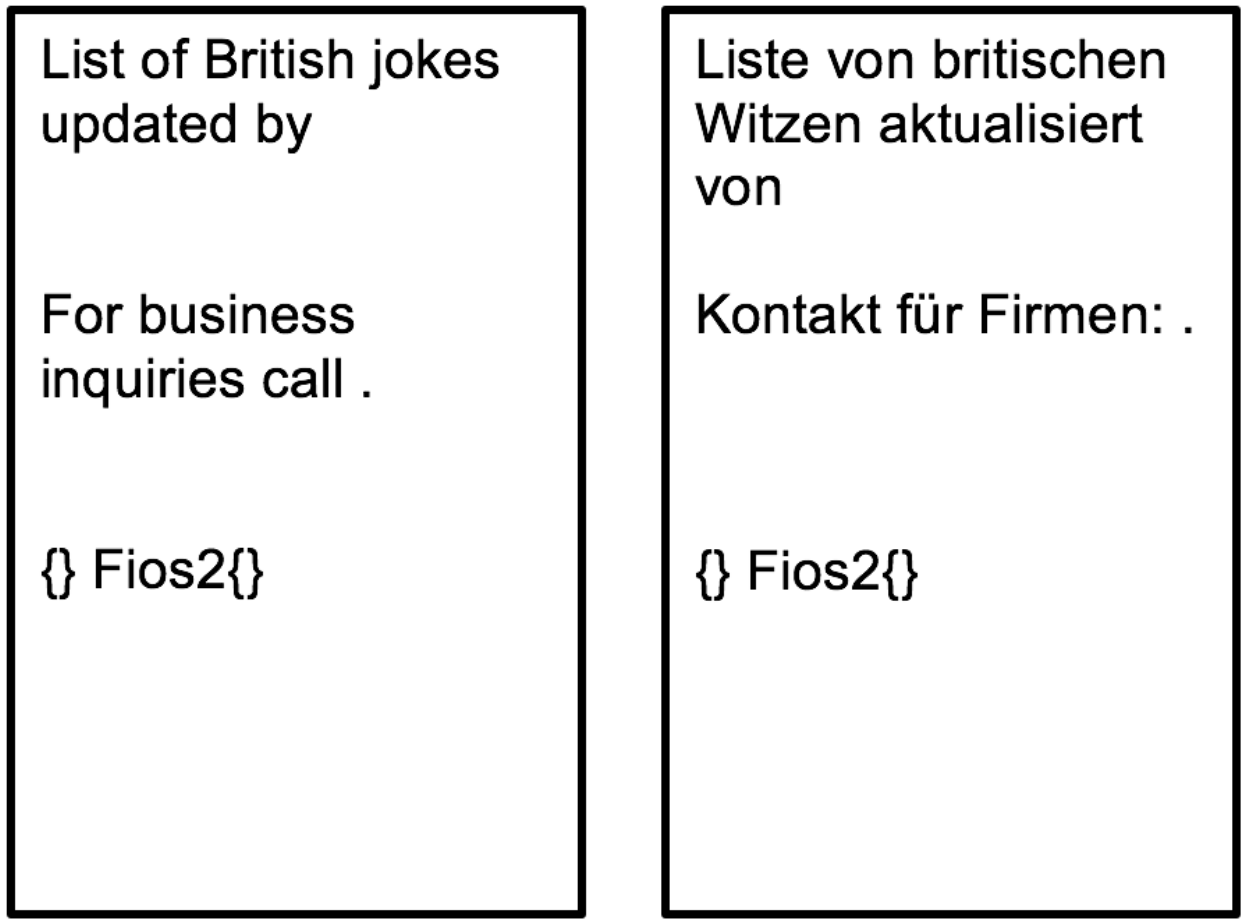
<bpt id="1" ctype="bold">{}</bpt>Fios2<ept id="1">{}</ept>

Liste von britischen Witzen aktualisiert von rowan@kinson.com

Kontakt für Firmen: +4140356 12 25.

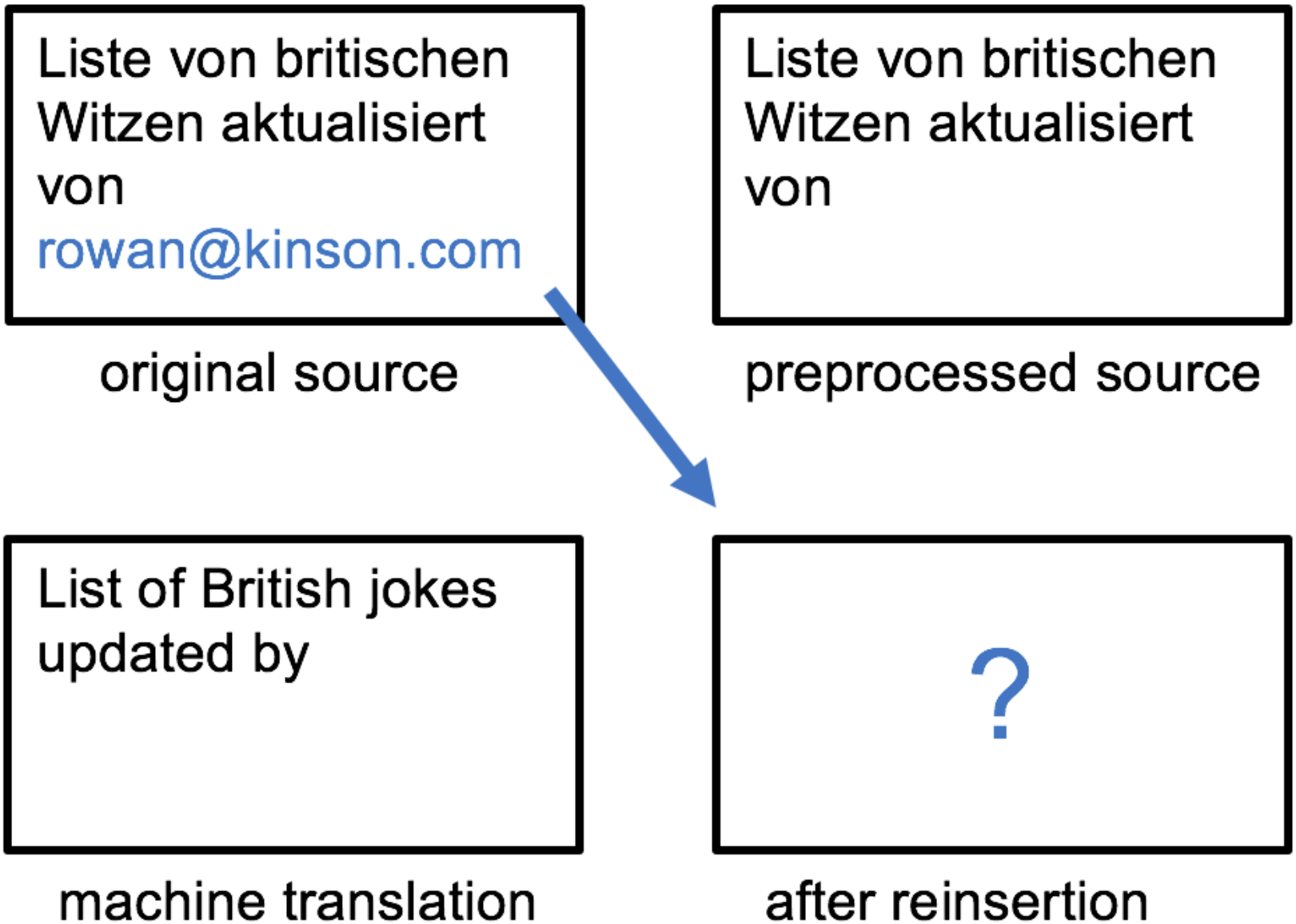
<bpt id="1" ctype="bold">{}</bpt>Fios2<ept id="1">{}</ept>

How Reinsertion Works

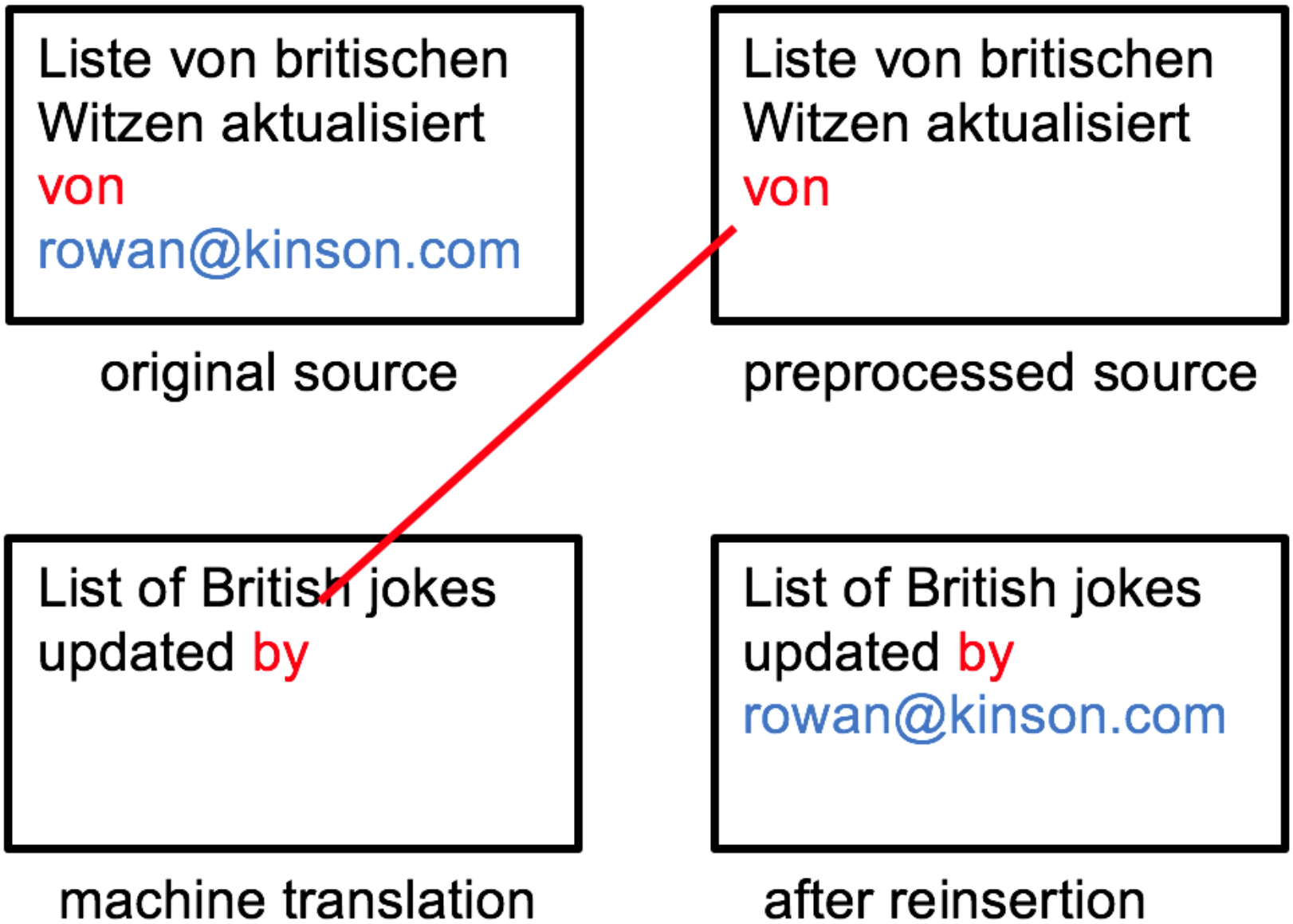


This is what the final training corpus looks like!

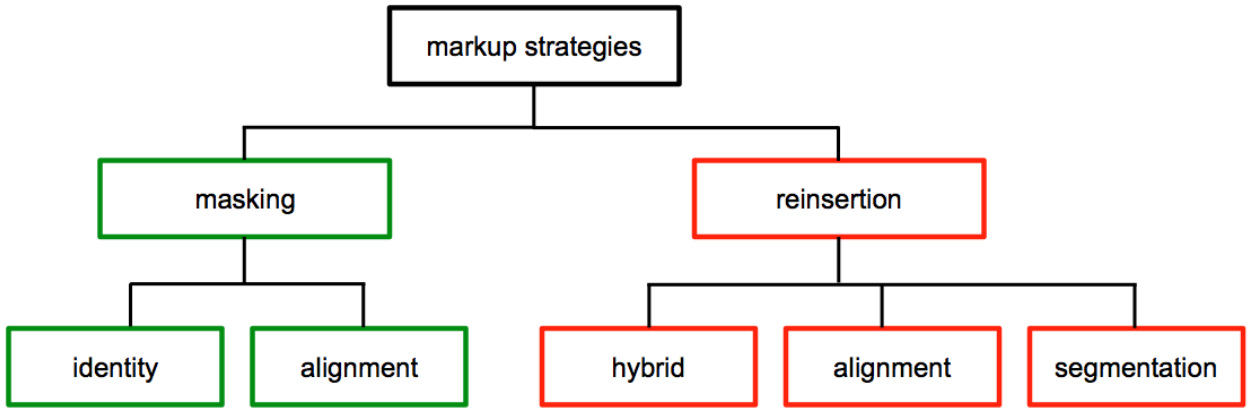
Undoing masking or markup removal



Using word alignment or similar correspondence



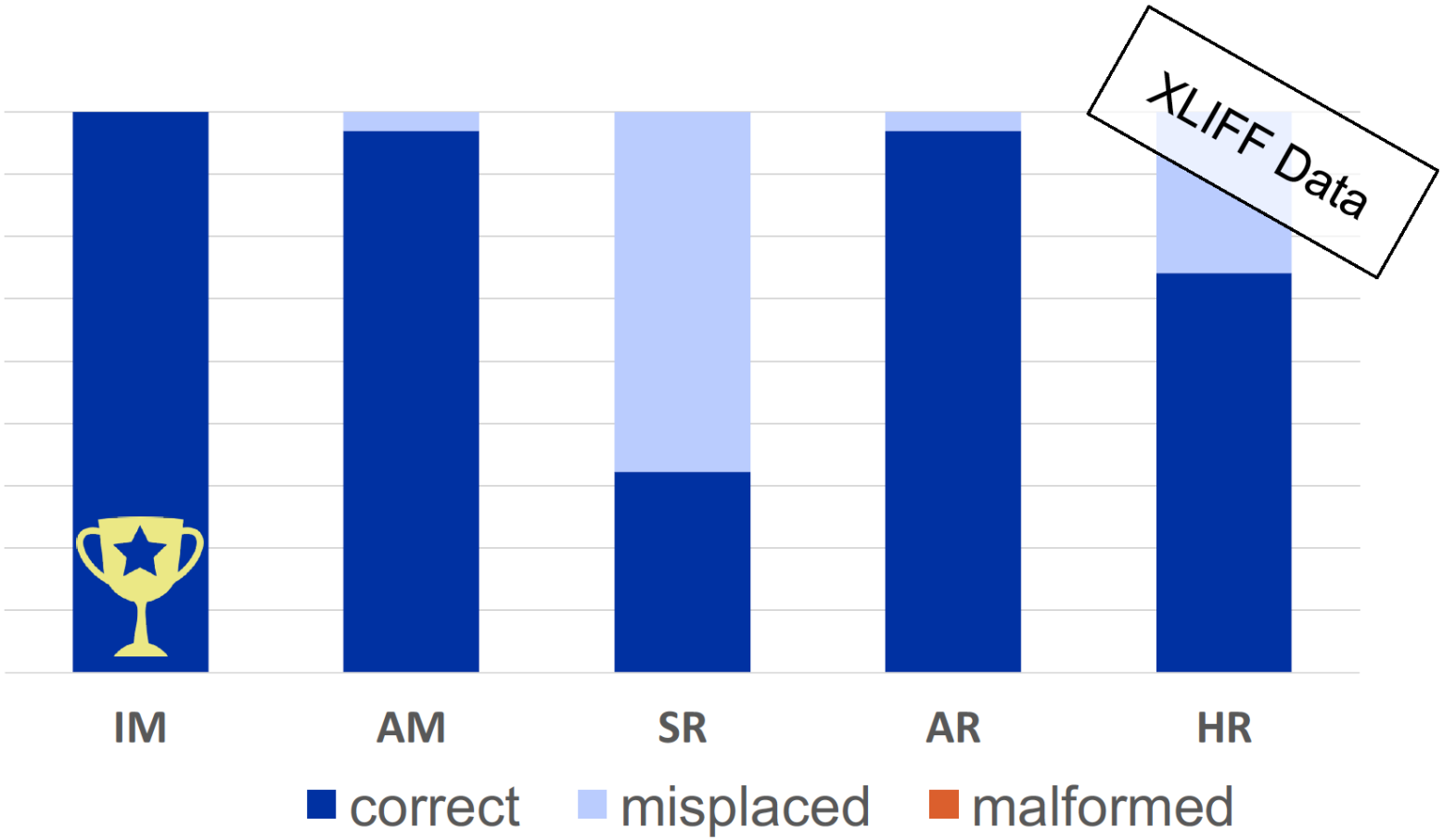
Experiments in a nutshell



- test 5 markup handling methods
- data sets where markup is abundant
- train standard Moses SMT systems, vary only the markup handling method
- manual evaluation of performance:

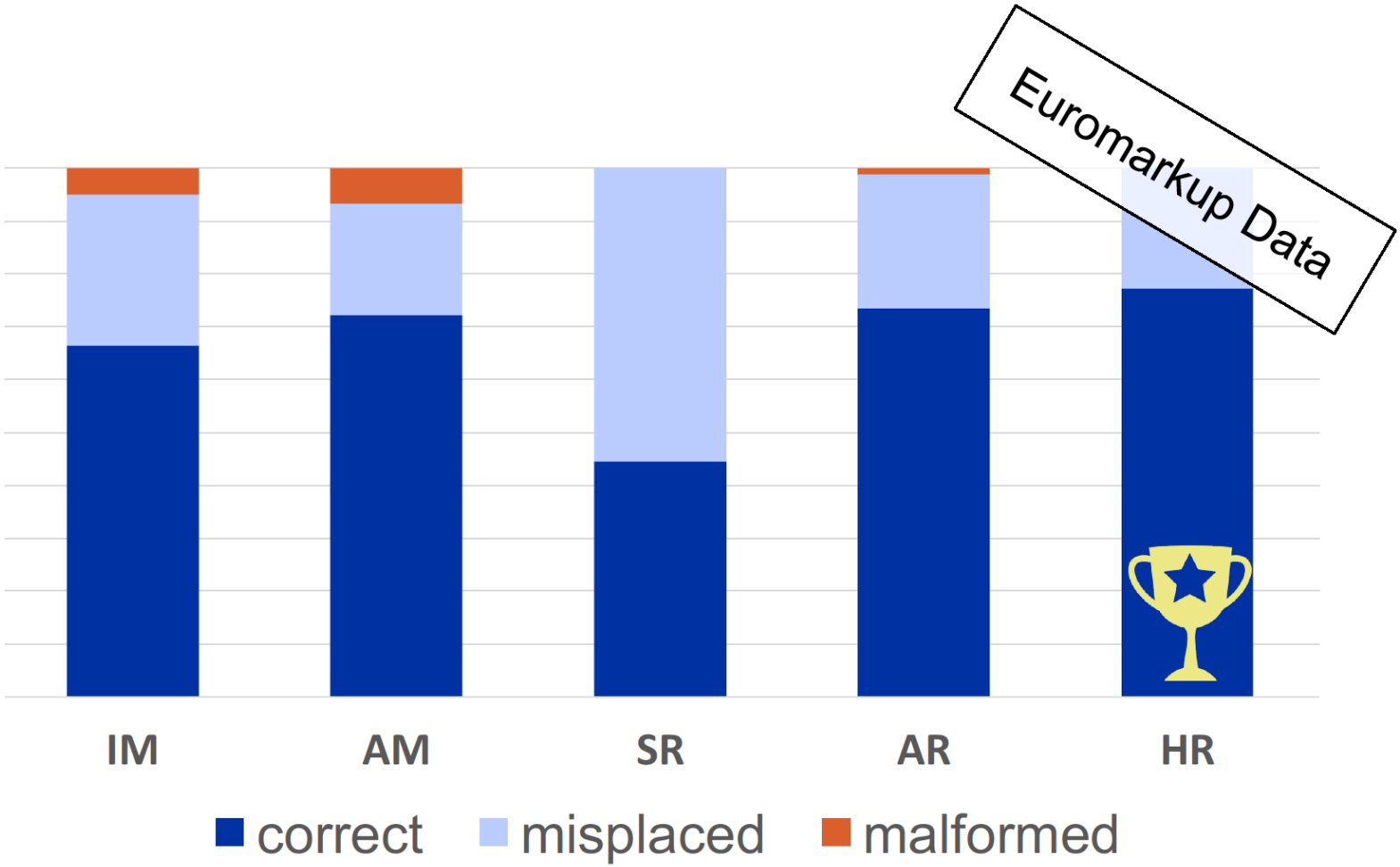
correct misplaced malformed

Manual evaluation of 568 tags in XLIFF data



IM = identity masking, AM = alignment masking, SR = segmentation reinsertion, AR = alignment reinsertion, HR = hybrid reinsertion.

Manual evaluation of 584 tags in Euromarkup data



IM = identity masking, AM = alignment masking, SR = segmentation reinsertion, AR = alignment reinsertion, HR = hybrid reinsertion.

Wrapping up

The main contributions of my thesis are:

- a survey of existing solutions for markup handling in SMT
- implementations of novel and existing methods in a unified framework that is available for free
- recommendations regarding the choice markup strategy

(... and other things I could not mention in all brevity :( )

Calling the mtrain API example

```
>>> from mtrain.preprocessing.reinsertion import
Reinserter
>>> reinserter = Reinserter('alignment')
>>> source_segment = 'Hello <g id="1"
ctype="x-bold;"> World ! </g>'
# markup removal, then translation...
>>> translated_segment = 'Hallo Welt !'
>>> alignment = [(0,0),(1,1),(2,2)]
>>> reinserter.reinsert(source_segment,
                        translated_segment,
                        alignment)
'Hallo <g ctype="x-bold;" id="1"> Welt ! </g>'
```

Thanks for listening!

First question for a lively discussion:
Yeah, but does this work for neural machine translation?

Bibliography I

Du, J., Roturier, J., and Way, A. (2010). TMX markup: a challenge when adapting smt to the localisation environment. In *EAMT - 14th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Hudík, T. and Ruopp, A. (2011). The integration of Moses into localization industry. In *15th Annual Conference of the EAMT*, pages 47–53.

Joanis, E., Stewart, D., Larkin, S., and Kuhn, R. (2013). Transferring markup tags in statistical machine translation: A two-stream approach. In O'Brien, S., Simard, M., and Specia, L., editors, *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, pages 73–81.

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine translation*, 25(3):197–215.

Parra Escartín, C. and Arcedillo, M. (2015). Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings. In *Proceedings of MT Summit XV*, pages 131–144. Association for Machine Translation in the Americas.

Tezcan, A. and Vandeghinste, V. (2011). SMT-CAT integration in a Technical Domain: Handling XML Markup Using Pre & Post-processing Methods. *Proceedings of EAMT 2011*.

Bibliography II

Zhechev, V. and van Genabith, J. (2010). Seeding statistical machine translation with translation memory output through tree-based structural alignment. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 43–51, Beijing, China. Coling 2010 Organizing Committee.