Crossing the Border Twice: Reimporting Prepositions to Alleviate L1-Specific Transfer Errors

Johannes Graën Gerold Schneider

Institute of Computational Linguistics
University of Zurich, Switzerland

22nd May, 2017

Outline

Motivation

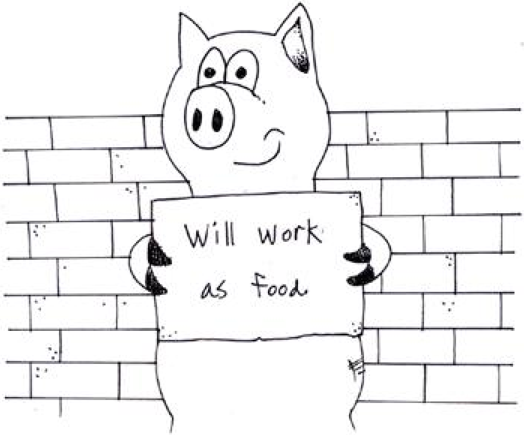
Corpus Material

Methods

Evaluations

Conclusions

Prepositions are important



Due to his grammar mistake, Wilbur found a position. It just wasn't the one he wanted.

Learner errors involving prepositions

ICLE

Has anybody any time stopped to think on the price that such advances have costed to humanity?

FCE

We can't imagine to live without it anymore because we are so dependent of it.

NICT

So I complain of him and ordered to take it back to me.

Verb-Preposition Constructions (VPC) and Adjective-Preposition Constructions (APC)

- VPC are difficult to acquire for language learners (Gilquin, Granger, et al. 2011, pp. 59–60).
- Phrasal verbs are "one of the most notoriously challenging aspects of English language instruction" (Gardner and Davies 2007, p. 339).
- We include APC as they are often similarly difficult to acquire for learners of English.
- In the CoNLL shared tasks for grammatical error correction, prepositional errors were the third most frequent error type at 5 to 9 % of all errors.

Background

VPC/APC are difficult for L2 language learners. Thus methods and tools for language learners are needed.

Schneider and Gilquin (2016) use & evaluate collocations to detect non-standard VPC: expected (E) collocational strength in Learner English (ICLE) compared to the observed (O) collocational strength in native English (from BNC):

$$\text{O/E-ratio} = \frac{\text{O/E(ICLE)}}{\text{O/E(BNC)}}$$

$$\text{t-ratio} = \frac{\text{t-score(ICLE)}}{\text{t-score(BNC)}}$$

Example: t-score ratio

T ratio	VERB	PREP	F	T(ICLE)	T(BNC)	COMMENT
5.9820	impose	to	10	5336.86	892.15	instead of *impose on*
3.5860	replace	to	3	1168.35	325.81	instead of *replaced by*
2.1133	accuse	for	8	5143.81	2433.98	instead of *accuse of*
2.0275	addict	on	4	3431.99	1692.68	instead of *addict to*
1.4296	better	than	87	17920.70	12535.47	
1.3929	alarm	of	2	2691.03	1932.01	instead of *alarm about*
1.3322	handicap	after	30	10530.89	7905.03	
1.2812	better	for	59	14564.98	11367.88	
1.2074	diverse	by	2	2690.71	2228.48	instead of *different according to*
1.1541	discuss	about	43	12421.43	10762.54	instead of *discuss sth.*
0.9322	consist	on	13	6290.72	6748.02	instead of *consist of*

⋮

Outline

Motivation

Corpus Material

Methods

Evaluations

Conclusions

Source

Europarl (version 7)

- Comprises transcript of the European Parliament sittings
- Contains numerous errors
- Has originally been compiled for training SMT systems
- Provides (reliable) alignment at the level of individual sittings

[1] http://pub.cl.uzh.ch/purl/costep

Source

Europarl (version 7)

- Comprises transcript of the European Parliament sittings
- Contains numerous errors
- Has originally been compiled for training SMT systems
- Provides (reliable) alignment at the level of individual sittings

CoStEP (Corrected & Structured Europarl Corpus; (Graën, Batinic, and Volk 2014))[1]

- Bases on the Europarl corpus
- Has undergone extensive cleaning
- Comprehends ca. 87 % of the original corpus material
- Provides alignment of speaker turns and additional speaker information

[1] http://pub.cl.uzh.ch/purl/costep

Our Corpus

Version 6

- 136,298 speaker turns from **CoStEP** in six languages (English, Finnish, French, German, Italian and Spanish) plus Polish whenever available (10 to 40 million tokens)
- Tokenization with our own multilingual tokenizer **Cutter**;[2] sentence segmentation based on tokenization tags
- Part-of-speech tagging and lemmatization with the **TreeTagger** and its featured language models
- Tag mapping to universal part-of-speech tags
- Dependency parsing with **MaltParser**
- Pairwise sentence alignment with **hunalign** and word alignment with the **Berkeley Aligner**

[2] http://pub.cl.uzh.ch/purl/cutter

Outline

Motivation

Corpus Material

Methods

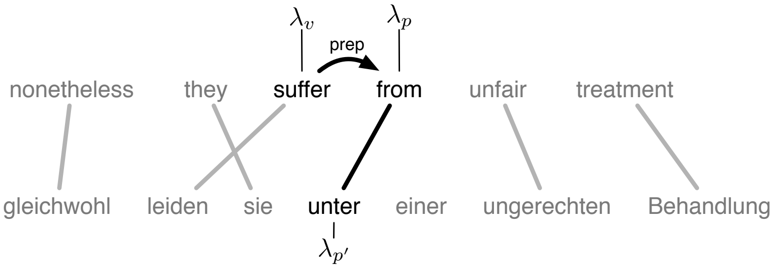
Evaluations

Conclusions

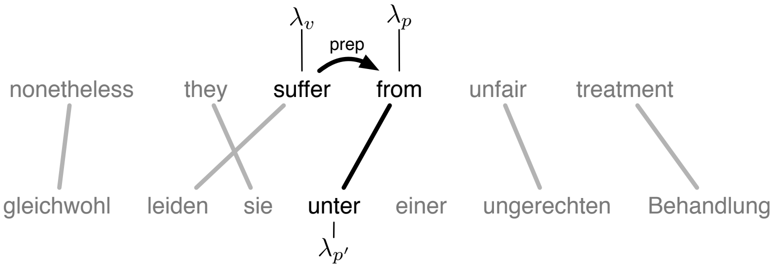
Lemma distribution matrix

- Based on word alignment and lemmatization.
- Reflects the probability of a lemma λ_s in the source language to be aligned with a lemma λ_t in the target language: $a(\lambda_t|\lambda_s)$
- The probabilities of all possible lemmas λ_i in the target language (i.e. the elements of the entire corresponding row) sum up to 1 by definition.

A verb, its preposition and the translated preposition



A verb, its preposition and the translated preposition



- λ_v – the verb (or adjective) lemma
- λ_p – the corresponding preposition
- $\lambda_{p'}$ – the translated preposition

Calculating distributions

- How often does the preposition λ_p appear with the verb λ_v?
- $f_V(\text{consist}, \text{of}) = 1146$
- $p_V(\text{of}|\text{consist}) = 82.7\,\%$
- How frequent is the translated preposition $\lambda_{p'}$ in language γ given the VPC (λ_v,λ_p)?
- $f_{V'}(\text{consist}, \text{of}, \text{german}, \text{aus}) = 121$
- $f_{V'}(\text{consist}, \text{of}, \text{german}, \text{von}) = 65$
- $f_{V'}(\text{consist}, \text{of}, \text{german}, \text{in}) = 38$
- ...

Calculating the backtranslation score and ratio

- Multiply the frequencies $f_{V'}$ of each translated preposition $\lambda_{p'}$ with the corresponding row of the lemma distribution matrix: $f_{V'}(\lambda_v, \lambda_p, \gamma, \lambda_{p'}) \times \big(a(\lambda_1|\lambda_{p'}), \ldots, (\lambda_n|\lambda_{p'})\big)$
- Sum up the columns (i.e. English lemma vectors) of the resulting rows to obtain the backtranslation scores (BTS)
- To attain the normalized backtranslation ratio (BTR), every element in the vector is divided by the BTS of the 'correct' preposition $(\lambda_{p''} = \lambda_p)$

Example: backtranslation via German

λ_v	λ_p	$\lambda_{p''}$	BTS	BTR
suffer	from	under	102.512	2.51
suffer	from	of	100.036	2.46
suffer	from	in	78.559	1.93
suffer	from	by	51.188	1.25
suffer	from	on	46.534	1.14
suffer	from	**from**	**40.966**	**1.00**
suffer	from	with	36.322	0.89
suffer	from	among	27.927	0.68
suffer	from	at	15.791	0.39
suffer	from	amongst	11.207	0.28

⋮

Outline

Motivation

Corpus Material

Methods

Evaluations

Conclusions

Evaluations

1. Do the expected errors occur in Learner corpora?
   - We consider those items that occur in each of the 5 language-specific lists as generally hard to learn. P $= 72\,\%$
   - OK?: is non-semantic prep; I: in ICLE; N: in NICT; F: in FCE
2. Can the errors be corrected?
   - We can correct $79\,\%$, upper bound is $96\,\%$.
   - Evaluation based on the errors found in ICLE by Schneider and Gilquin (2016)
   - CORR: suggested correction; MATCH?: is suggestion correct?
   - *obj* or *PP* as first decision: *obj* if VPC $< 33\,\%$

VERB/ADJ	PREP	OK?	I	N	F
aim	at	yes	+		
arrive	at	yes	+	+	+
benefit	from	yes	+		
breathe	into	?		*n/a*	
channel	into	yes		*n/a*	
complain	about	yes	+	+	+
compliment	on	yes			
convert	into	yes		*n/a*	
depend	on	yes	+		+
		⋮			
talk	about	yes	+	+	+
target	at	yes	+		
throw	into	?		*n/a*	
transform	into	?		*n/a*	
translate	into	?		*n/a*	
transpose	into	?		*n/a*	
wait	for	yes	+	+	+
worry	about	yes			+
Total		34/10/3		23/31	

VERB/ADJ	PREP	CORR	MATCH?
accuse	for	of	yes
addict	on	to	yes
alarm	of	at	yes
apply	into	to	yes
assist	to	*obj*	yes
assure	to	*obj*	yes
aspire	for	to	yes
attack	against	*obj*	yes
aware	about	of	yes
	⋮		
relate	with	to	yes
replace	to	by	no
resist	to	*obj*	yes
select	among	from	no
separate	between	*n/a*	no
study	about	*obj*	yes
understand	towards	*obj*	yes
view	upon	on	no
Total			38/48

Outline

Motivation

Corpus Material

Methods

Evaluations

Conclusions

Conclusion

- We have employed word alignment in a large parallel corpus to identify potentially difficult VPC/APC, without needing annotated resources or learner corpora.
- We offer language-specific VPC/APC lists ranked by a combined measure of difficulty and frequency.
- Intersecting these lists reports generally difficult VPC/APC.[3]
- Romance languages, as expected, exhibit a larger overlap of combinations than other languages.
- We have evaluated our method in two ways
  - How many of the VPC/APC items in our lists are found in Learner language?
  - How many of the suggested corrections are appropriate?

[3] http://pub.cl.uzh.ch/purl/reimporting_prepositions

Outlook

- We intend to extend our approach to further languages and other constructions in future research.
- Tuning our alignment approach with gold standard data, such as thresholds and filters, and use further corpora from different genres.
- Distinguish complements from adjuncts.
- Improve alignment and parsing.
- Respect the translation direction and the influence of fixed idioms.
- Recruit example sentences in which the difficult VPC occur.
- Involve learners and language centres in the evaluation and teaching.

References I

Dee Gardner and Mark Davies (2007). “Pointing Out Frequent Phrasal Verbs: A Corpus-Based Analysis”. In: *TESOL quarterly* 41.2, pp. 339–359

Gaëtanelle Gilquin, Sylviane Granger, et al. (2011). “From EFL to ESL: evidence from the International Corpus of Learner English”. In: *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*, pp. 55–78

Johannes Graën, Dolores Batinic, and Martin Volk (2014). “Cleaning the Europarl Corpus for Linguistic Applications”. In: *Proceedings of the Conference on Natural Language Processing (KONVENS).* (Hildesheim). Stiftung Universität Hildesheim, pp. 222–227

Gerold Schneider and Gaëtanelle Gilquin (2016). “Detecting Innovations in a Parsed Corpus of Learner English”. In: *International Journal of Learner Corpus Research* 2.2