The MuCoW test suite at WMT 2019: Automatically harvested **mu**ltilingual **co**ntrastive **w**ord sense disambiguation test sets for machine translation

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS

Alessandro Raganato, Yves Scherrer, Jörg Tiedemann
https://github.com/Helsinki-NLP/MuCoW

What is MuCoW?

MuCoW is a language-independent method for automatically building a ***lexical ambiguity benchmark*** for machine translation based on contrastive translation pairs.

MuCoW focuses on lexical ambiguity:

Words of the source language that have multiple translations in the target language, representing different meanings.

MuCoW comes in two variants:

The ***scoring variant*** covers 11 language pairs with a total of almost 240 000 sentence pairs.

The ***translation variant*** covers 9 language pairs with a total of 15 600 sentences.

The tools

BabelNet is a multilingual encyclopedic dictionary made up of about 16 million entries, called Babel synsets. Each Babel synset represents a meaning and contains all the synonyms which express that meaning in a range of different languages. https://babelnet.org

SW2V is a neural model that learns word and synset embeddings in a shared vector space. http://lcl.uniroma1.it/sw2v

OPUS is a collection of translated texts from the web. http://opus.nlpl.eu

Eflomal is a fast and accurate word alignment tool that uses Gibbs sampling with a Bayesian extension of the IBM models. https://github.com/robertostling/eflomal

Step 1

Identify ambiguous source words and their translations

Apply the ***Eflomal*** word alignment tool on a collection of parallel corpora from ***OPUS***:
Books, EU Bookshop, Europarl, MultiUN, News-Commentary, OpenSubtitles, SETIMES, Tatoeba, TED

Example: English words aligned to German *Eingabe*

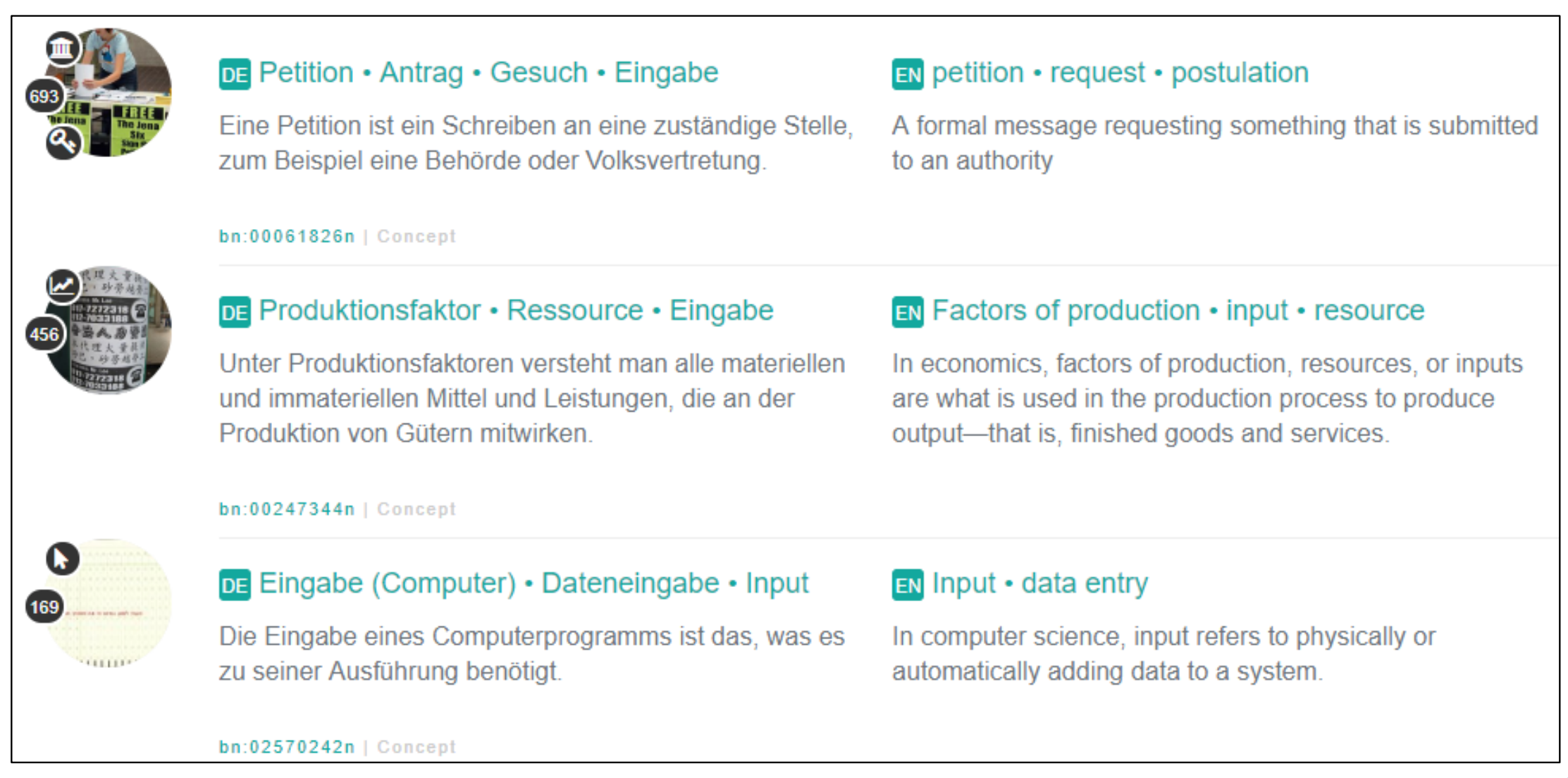
177 input	26 documents	9 system
50 typing	21 petition	8 entered
29 entering	17 data	8 command
28 entry	14 submission	7 display
27 loading	13 the	7 to
26 enter	11 inputting	…

Step 2a

Cluster target words via BabelNet

Query ***BabelNet*** with each ambiguous source word.

Take the intersection of the alignment-inferred target words and the BabelNet-inferred target words.



693

DE Petition • Antrag • Gesuch • Eingabe
Eine Petition ist ein Schreiben an eine zuständige Stelle, zum Beispiel eine Behörde oder Volksvertretung.

EN petition • request • postulation
A formal message requesting something that is submitted to an authority

bn:00061826n | Concept

456

DE Produktionsfaktor • Ressource • Eingabe
Unter Produktionsfaktoren versteht man alle materiellen und immateriellen Mittel und Leistungen, die an der Produktion von Gütern mitwirken.

EN Factors of production • input • resource
In economics, factors of production, resources, or inputs are what is used in the production process to produce output—that is, finished goods and services.

bn:00247344n | Concept

169

DE Eingabe (Computer) • Dateneingabe • Input
Die Eingabe eines Computerprogramms ist das, was es zu seiner Ausführung benötigt.

EN Input • data entry
In computer science, input refers to physically or automatically adding data to a system.

bn:02570242n | Concept

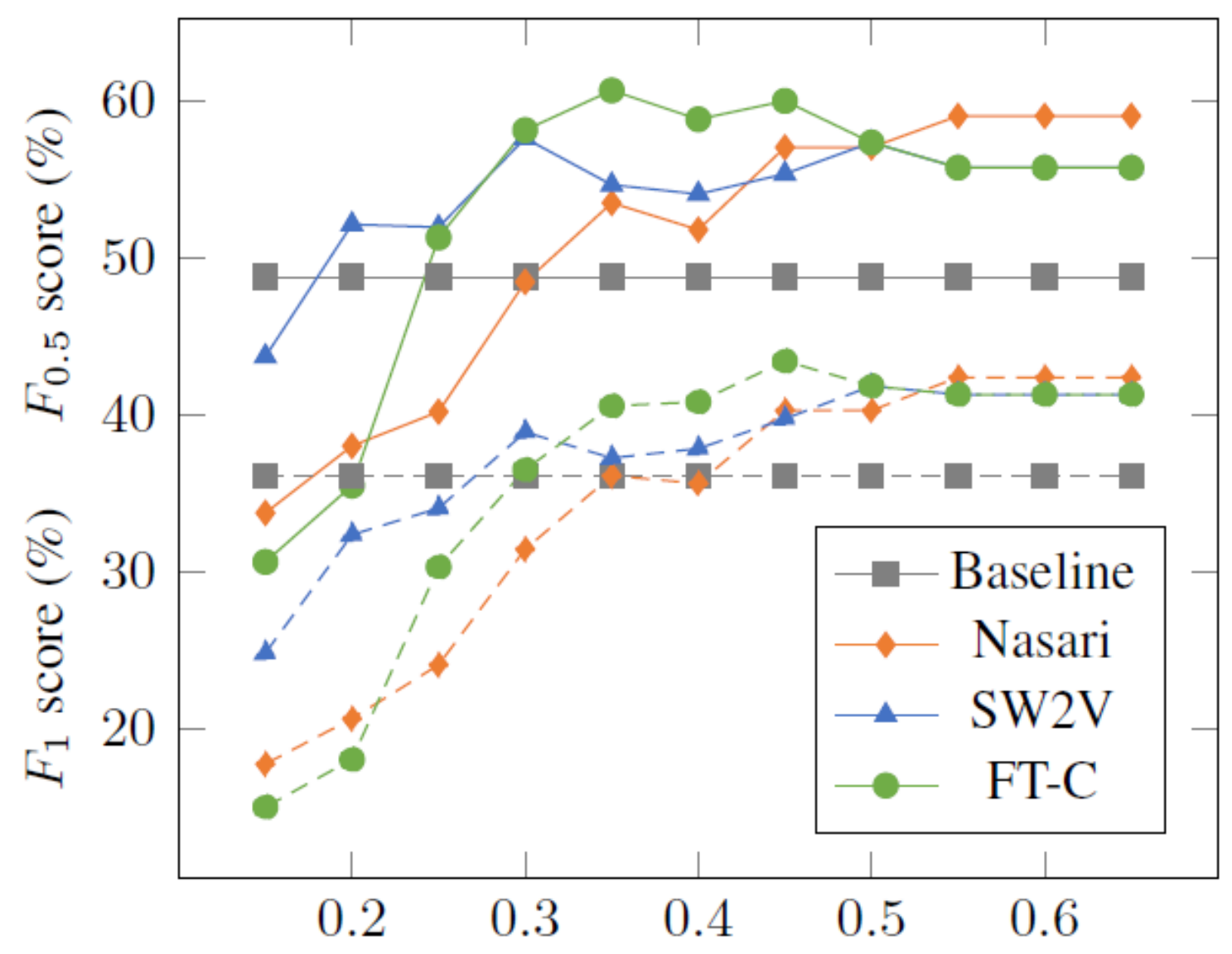
Step 2b

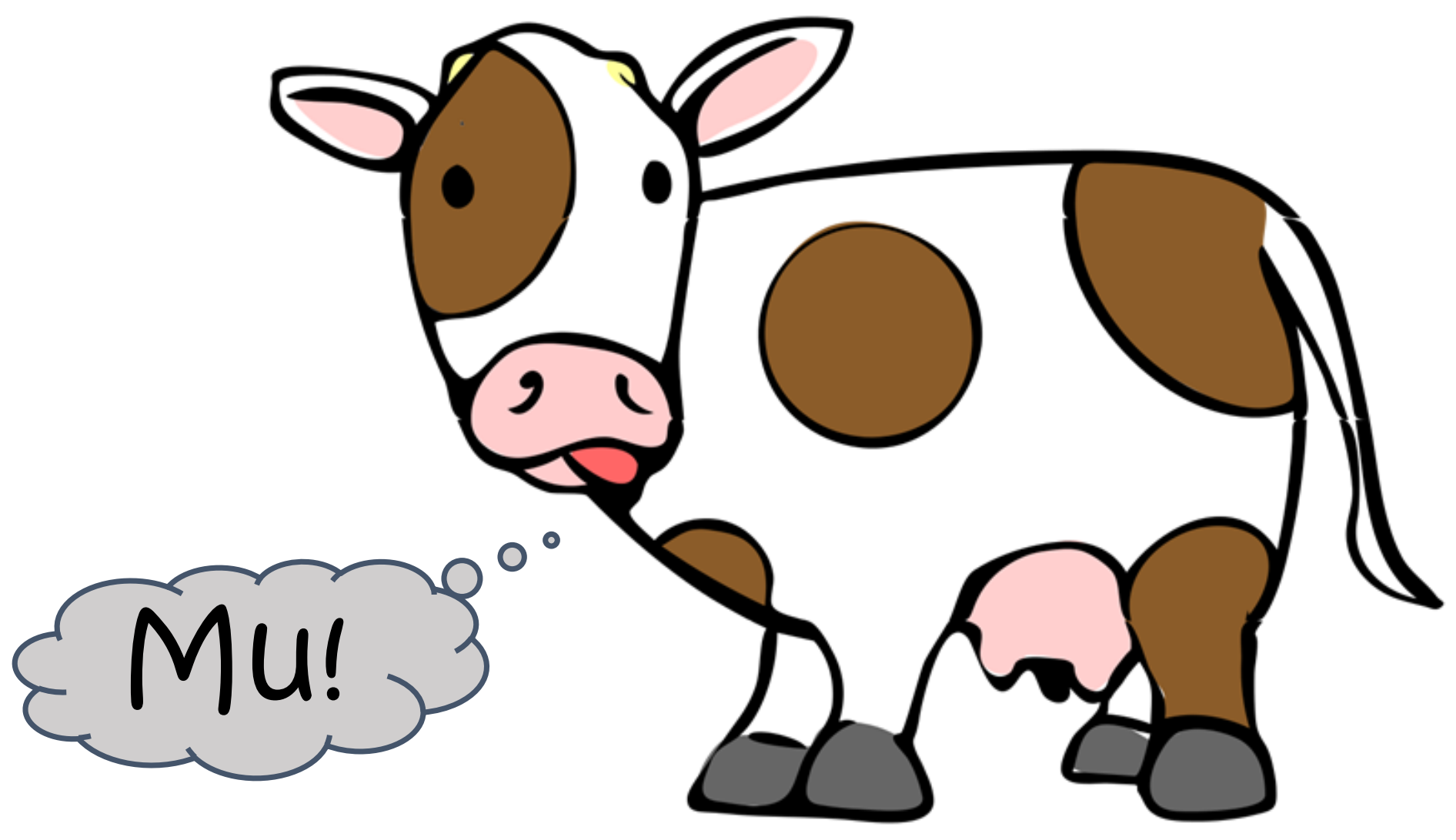
Refine sense clusters with sense embeddings

Associate each Babel synset with its ***SW2V*** embedding.

Compute pairwise cosine similarities between synsets.

Merge them if their similarity is higher than threshold γ.





Example containing **ambiguous word**	Correct translations	Incorrect translations
It occurred to me that my **watch** might be broken. I hope you didn't get distracted during your **watch**.	Armbanduhr, Uhr *Wache*	*Wache* Armbanduhr, Uhr
In winter, the dry leaves fly around in the **air**. He remained silent for a moment, with a thoughtful but contented **air**.	Luft, Luftraum, Aura Miene, Ausdruck	Miene, Ausdruck Luft, Luftraum, Aura
Harry had to back out of the competition because of a broken **arm**. So does the cop who left his side arm in a subway bathroom.	Arm *Waffe*	*Waffe* Arm
Drain the pasta and return the pasta to the **pot**. Where did those idiots get all of this **pot** anyhow?	Blumentopf, Kochtopf, Topf, Nachttopf *Marihuana, Gras*	*Marihuana, Gras* Blumentopf, Kochtopf, Topf, Nachttopf

Findings

Research systems perform poorly on out-of-domain synsets, whereas online systems are more robust.

From-English directions show higher overall precision than to-English directions: less reliable encoder representations for morphologically rich languages?

Step 3 – Scoring variant

Create contrastive sentence pairs

Extract sentence pairs from the parallel corpora and group them by source word and target word sense, using the synset lexicon built in Step 2b.

For each extracted sentence pair, a contrastive sentence pair is produced by replacing the target word in the target sentence by another lexicalisation from a different synset.

Statistics:

	Corpus	Lexicon			Test suite
Language pair	Sentence pairs	Source words	Target synsets	Target words	Sentence pairs
CS–EN	44M	107	223	412	11470
DE–EN	35M	259	548	1086	33077
ES–EN	81M	515	1090	2398	72295
ET–EN	14M	34	68	89	2500
FI–EN	31M	176	367	610	16326
FR–EN	68M	456	963	2152	64369
LT–EN	2.5M	10	20	31	922
LV–EN	1.6M	5	10	12	318
RO–EN	52M	129	263	496	14258
RU–EN	38M	113	234	396	12378
TR–EN	46M	107	220	420	11795

Evaluation results (accuracy):

Lg. pair	Model	ContraWSD	MuCoW	BLEU
DE–EN	LSTM	77.55	60.50	30.3
	Transformer	86.42	66.98	33.3
	Nematus	86.72	68.80	35.1
CS–EN	Nematus		78.77	30.9
RO–EN	Nematus		62.86	33.3
RU–EN	Nematus		72.36	30.8
TR–EN	Nematus		62.69	20.1

Step 3 – Translation variant

Extract sense-annotated sentences

Extract sentence pairs from the parallel corpora and group them by source word and target word sense, using the synset lexicon built in Step 2b.

Associate the source sentences with a set of correct lexicalizations and a set of incorrect lexicalizations.

Apply additional filters

Part-of-speech filtering: only keep sentence pairs in which both the source and target words are tagged as NOUNs.

Corpus filtering: exclude sentences stemming from one of the WMT training corpora.

Domain annotation: split the senses into in-domain (<= 50% OpenSubtitles) and out-of-domain (> 50% OpenSubtitles).

Statistics:

Language pair	Source words	Target synsets	In-dom synsets	Out-dom synsets	Sen-tences
DE–EN	217	461	329	132	4268
FI–EN	109	231	91	140	2117
LT–EN	6	12	5	7	99
RU–EN	67	138	59	79	1223
EN–CS	98	200	29	171	1843
EN–DE	176	362	220	142	3337
EN–FI	48	97	22	75	830
EN–LT	4	8	3	5	69
EN–RU	97	199	40	163	1814

WMT test suite results (Top 3 per direction)

	In-domain synsets			Out-of-domain synsets			All synsets			Human
Submission	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	Rank
English–Czech:										
CUNI-Trf-T2T-2018	96.76	84.75	90.36	**79.85**	**71.71**	**75.56**	**82.77**	74.01	**78.15**	2
CUNI-Trf-T2T-2019	95.60	85.66	90.36	79.58	71.57	75.36	82.38	**74.04**	77.99	3
CUNI-DocTrf-T2T	95.60	85.66	90.36	79.58	71.57	75.36	82.38	**74.04**	77.99	1
German–English:										
Facebook_FAIR	**80.78**	**85.80**	**83.21**	**52.77**	**72.56**	**61.10**	**73.55**	**82.99**	**77.99**	1
online-B	77.88	83.81	80.73	45.50	66.51	54.04	69.58	80.31	74.56	4
online-G	77.62	83.76	80.57	45.62	65.43	53.76	69.48	80.02	74.38	14
English–German:										
Facebook_FAIR	**83.43**	76.99	**80.08**	**56.29**	**55.10**	**55.69**	**74.48**	**70.05**	**72.19**	1
Microsoft-sentence-level	83.18	**77.14**	80.05	52.81	51.92	52.36	73.31	69.27	71.23	11
online-B	83.37	74.78	78.85	51.92	50.66	51.28	73.04	67.30	70.05	10
Finnish–English:										
online-G	78.00	84.17	80.97	**71.47**	81.65	**76.22**	**74.14**	**82.71**	**78.19**	8
online-Y	79.30	82.89	81.05	63.40	**81.73**	71.41	69.78	82.25	75.51	2
GTCOM-Primary	81.87	**84.81**	**83.31**	57.28	77.64	65.92	67.36	81.05	73.57	3
English–Finnish:										
online-G	93.71	75.25	83.47	**80.62**	**68.54**	**74.09**	**84.01**	**70.36**	**76.58**	6
online-Y	94.74	72.00	81.82	75.06	66.08	70.28	80.03	67.75	73.38	3
MSRA.NAO	**95.62**	**76.12**	**84.76**	68.47	66.60	67.52	75.44	69.42	72.31	2
Lithuanian–English:										
tilde-c-nmt							**80.41**	97.50	**88.14**	5
NEU							79.59	**98.73**	**88.14**	3
tilde-nc-nmt							79.38	97.47	87.50	2
English–Lithuanian:										
MSRA.MASS							78.69	**85.71**	**82.05**	2
online-B							79.31	80.70	80.00	8
tilde-nc-nmt							80.70	79.31	80.00	1
Russian–English:										
online-G	**92.15**	89.63	**90.87**	**66.95**	**80.87**	**73.26**	**78.57**	**85.38**	**81.84**	2
Facebook_FAIR	89.98	**89.80**	89.89	56.67	77.30	65.40	72.12	84.07	77.64	1
online-B	89.55	87.58	88.55	56.41	74.07	64.04	71.81	81.34	76.28	4
English–Russian:										
online-G	**95.56**	89.58	92.47	**75.11**	**74.85**	**74.98**	**80.05**	**78.58**	**79.31**	3
Facebook_FAIR	95.49	88.28	91.75	67.68	71.54	69.56	74.40	76.01	75.20	1
online-B	95.08	91.10	93.05	62.12	69.05	65.40	70.31	75.16	72.66	4