Verb Sense Disambiguation in Machine Translation

Roman Sudarikov, Ondřej Dušek, Martin Holub, Ondřej Bojar, and **Vincent Kríž**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{sudarikov,odusek,holub,bojar,kriz}@ufal.mff.cuni.cz

Abstract

We describe experiments in Machine Translation using word sense disambiguation (WSD) information. This work focuses on WSD in verbs, based on two different approaches – verbal patterns based on corpus pattern analysis and verbal word senses from valency frames. We evaluate several options of using verb senses in the source-language sentences as an additional factor for the Moses statistical machine translation system. Our results show a statistically significant translation quality improvement in terms of the BLEU metric for the valency frames approach, but in manual evaluation, both WSD methods bring improvements.

1 Introduction

The possibility of using word sense disambiguation (WSD) systems in machine translation (MT) has recently been investigated in several ways: Output of WSD systems has been incorporated into MT to improve translation quality — at the decoding step of a phrase-based statistical machine translation (PB-SMT) system (Chan et al., 2007) or as contextual features in maximum entropy (MaxEnt) models (Neale et al., 2015) and (Neale et al., 2016). In addition, WSD has also been used in MT evaluation, for example in METEOR (Apidianaki et al., 2015). These works indicate that WSD can be beneficial to different MT tasks, in case of using senses as contextual features for MaxEnt models Neale et al. (2016) achieve statistically significant improvement over the baseline for English-to-Portuguese translation. And Apidianaki et al. (2015) report that usage of WSD can establish better sense correspondences and improve its correlation with human judgments of translation quality.

In this research, we have investigated the possibilities of integrating two different approaches to verbal WSD into a PB-SMT system – verb patterns based on corpus pattern analysis (CPA) and verbal word senses in valency frames. The focus on verbs was motivated by the ideas that verbs carry a crucial part of the meaning of the sentence (Healy and Miller, 1970) and thus accurate translation of the verb is critical for the understanding of the translation. Therefore, improvement of the translation of verbs can lead to overall increase of the translation quality. Therefore, improvement of the translation of verbs can lead to an overall increase of translation quality. The outputs of automatic verb sense disambiguation systems using both CPA and valency frames were integrated into Moses statistical machine translation system(Koehn et al., 2007). Both kinds of verb senses were added as additional factors (Koehn and Hoang, 2007). Section 4.1 shows that we obtain statistically significant improvement in terms of BLEU scores (Papineni et al., 2002) and manual evaluation of translations validated that.

The novelty of this work lies not only in our focus only on verbs senses, but also in the fact that we are comparing the impact of two WSD approaches on the statistical machine translation.

The following Section 2 describes the initial setup of our experiments. Section 3 and Section 4 depict the idea behind corpus pattern analysis and verb valency frames representations and show evaluation results of incorporation of these sense to phrase-based statistical machine translation. The next section (Section 5) is devoted to the discussion of results obtained during the evaluation. And finally Section 6 describes our plan of the future work.

Proceedings of the Sixth Workshop on Hybrid Approaches to Translation,
pages 42–50, Osaka, Japan, December 11, 2016.

2 Experiments setup

2.1 Dataset and MT system

For our experiments, we have used a subset of the Czech-English corpus CzEng 1.0 (Bojar et al., 2012); the respective numbers of sentences and tokens in each of training, development and test sets are shown in Table 1. For our experiments, 28 different English verbs were selected and automatically annotated with corpus pattern analysis senses, and 3,306 verbs annotated using valency frames. The subset has been selected to include verbs annotated with CPA, so the effect of WSD would be visible. All the experiments were carried out in the Eman experiment management system (Bojar and Tamchyna, 2013) using the Moses PB-SMT system (Koehn et al., 2007) as the core and minimum error rate training (MERT, (Och, 2003)) to optimize the decoder feature weights on the development set. The evaluation was performed using the BLEU score (Papineni et al., 2002), but the results of each setup were then thoroughly examined and verified using the MT-ComparEval system (Aranberri et al., 2016)[1].

Set	Number of sentences	Tokens CS	Tokens EN
Training	649,605	10,759,546	12,073,130
Development	10,115	187,478	167,788
Test	2,707	59,446	67,336

Table 1: Data set composition

2.2 MT configurations

As we have mentioned in Section 1 the main goal of the experiments was to explore whether verb senses as additional factors in the statistical MT system Moses can help in improving translation quality. The following configurations were tested:

- Form→Form – "vanilla" Moses setup, translating from surface word forms to target surface forms, including capitalization.

- Form+Sense→Form – two source factors (surface word form and verb sense ID, if applicable) are translated to the target-side word forms. This is technically identical to appending the verb sense ID to the source words.

- Form→Form+Tag – the source word form is translated to two factors on the target side: word form and morphological tag (part-of-speech tag with morphological categories of Czech, such as case, number, gender, or tense). This allows us to use an additional language model trained on morphological tags only. This setup is known to perform well for morphologically rich languages (Bojar, 2007) and thus was selected as a baseline for all comparisons.

- Form+Sense→Form+Tag – a combination of the two setups above: two source and two target-side factors, for better handling of source verb meaning and target morphological coherence.

- Form→Form+Tag + Form+Sense→Form+Tag – a combination of previous two models as two separate phrase tables.

For all configurations, we trained a 4-gram language model on word forms of the sentences from the training set. This LM was pruned: we discarded all singleton n-grams (apart from unigrams). In addition, for configurations which generated morphological tags, we used a 10-gram model LM over morphological tags to help maintain morphological coherence of the translation outputs. Again, we pruned all singleton n-grams with the exception of unigrams.

[1] `http://wmt.ufal.cz/`

Verb	No.	Pattern / Implicature
gleam	1	**[[Physical Object | Surface]] gleam [NO OBJ]** [[Surface]] of [[Physical Object]] reflects occasional flashes of light
gleam	2	**[[Light | Light Source]] gleam [NO OBJ]** [[Light Source]] emits an occasional flash of [[Light]]
gleam	3	{**eyes**} **gleam [NO OBJ] (with [[Emotion]])** {eyes} of [[Human]] shine, expressive of [[Emotion]]
wake	3	**[no object] [Human] wake** ({**up**}) **AdvTime**({**from**} {**nightmare** | **dream** | **sleep** | **reverie**}) ({**to**} **Eventuality)** the mind of [[Human]] returns at a particular [[Time]] to a state of full conscious awareness and alertness after sleep
wake	4	**pv [phrasal verb] [[Human 1] ^ [Sound] ^ [Event]] wake [[Human 2] ^ [Animal]]** ({**up**}) [[Human 1 — Sound — Event]] causes the mind of [[Human 2 — Animal]] to return to a state of full conscious awareness and alertness after sleep
wake	7	**[Anything] wake [Emotion]** ({**in**} **Human)** [[Anything]] causes [[Human]] to feel or become aware of [[Emotion]]
wake	9	**waking *** ({**up**}) [Human—Animal]'s returning to a state of full conscious awareness and alertness after sleep

Table 2: Example patterns defined for the verbs *gleam* and *wake*.

3 Verb patterns based on Corpus Pattern Analysis

Corpus Pattern Analysis (CPA) is a method of manual context-based lexical disambiguation of verbs (Hanks, 1994; Hanks, 2013). Verbs are supposed to have no meanings on their own; instead, meanings are triggered by the context. Hence, a CPA-based lexicon does not group the uses of a verb into senses but into syntagmatic usage patterns derived from the corpus findings. Such a CPA-based lexicon is the Pattern Lexicon of English Verbs (PDEV, (Hanks and Pustejovsky, 2005)). In contrast to the classical WSD, here the verb patterns are used as verb meaning representations. An example of a few patterns is given in Table 2.

Here we employ an automatic procedure for verb pattern recognition developed by Holub et al. (2012), which deals with 30 selected English verbs. In fact, their method uses 30 separate classifiers, one for each verb, trained on moderately sized manually annotated samples. They use the collection called VPS-30-En (Verb Pattern Sample, 30 English verbs) published by Cinková et al. (2012) as training data. VPS-30-En was designed as a small sample of PDEV, a pilot lexical resource of 30 English lexical verb entries enriched with semantically annotated corpus samples. The data describes regular contextual patterns of use of the selected verbs in the British National Corpus, version 3 (BNC, 2007).[2] The number of different patterns varies from 4 to 10 in most cases across the verbs, and the performance of Holub et al. (2012)'s automatic pattern recognition also differs verb from verb, ranging between 50% and 90% accuracy.

3.1 Experiments and evaluation

For the experiments with verb patterns based on CPA, we have explored all the configurations described in Section 2.2.

Table 3 shows the results of the best MERT run for each configuration. Multiple MERT runs evaluation was performed for Form→Form+Tag, Form+Sense→Form+Tag, and Form→Form+Tag + Form+Sense→Form+Tag using MultEval system (Clark et al., 2011) with Form→Form+Tag as the baseline system, and the results are shown in Table 4. We see that the average results of Form+Sense→Form+Tag are worse than the ones of Form→Form+Tag by 0.1% BLEU. MultEval aims to determine whether an experimental result has a statistically reliable difference for a give evaluation metric, using a stratified approximate randomization (AR) test. AR estimates the probability (p-value) that a measured difference in metric scores arose by chance by randomly exchanging sentences between the two systems. If there is no significant difference between the systems (i.e., the null hypothesis is true), then this shuffling should not change the computed metric score (Clark et al., 2011). While comparing

[2]Details about both selected verbs and training contexts can be found at `http://ufal.mff.cuni.cz/spr`.

Configuration	BLEU
Form→Form	24.26
Form+Sense→Form	24.15
Form+Sense→Form+Tag	25.01
Form→Form+Tag	25.11
Form→Form+Tag + Form+Sense→Form+Tag	25.27

Table 3: Evaluation results for corpus pattern analysis annotation, best MERT run

Form→Form+Tag and Form→Form+Tag + Form+Sense→Form+Tag, we see that p-value is 0.16, thus allowing us to claim, that these two systems don't differ one from another. The same test performed using METEOR and TER tests only confirms that (in case of TER having p-value=0.61).

Metric	**System**	**Avg**	$\bar{s}_{\text{sel}}$	s_{Test}	p**-value**
BLEU	Form→Form+Tag	25.0	0.9	0.1	-
	Form+Sense→Form+Tag	24.9	0.9	0.1	0.00
	Form→Form+Tag + Form+Sense→Form+Tag	25.0	0.9	0.1	0.16
METEOR	Form→Form+Tag	22.6	0.4	0.0	-
	Form+Sense→Form+Tag	22.5	0.4	0.0	0.00
	Form→Form+Tag + Form+Sense→Form+Tag	22.6	0.4	0.1	0.22
TER	Form→Form+Tag	62.2	0.7	0.2	-
	Form+Sense→Form+Tag	62.4	0.7	0.1	0.00
	Form→Form+Tag + Form+Sense→Form+Tag	62.2	0.7	0.2	0.61

Table 4: Multeval results for corpus pattern analysis, based on 36 MERT runs

We also performed a more detailed analysis with pairwise comparisons of the following configurations:

- Form→Form vs. Form+Sense→Form
- Form→Form+Tag vs. Form+Sense→Form+Tag
- Form→Form+Tag vs. Form→Form+Tag + Form+Sense→Form+Tag

3.1.1 Form→Form vs. Form+Sense→Form

The comparison provided by MT-ComparEval based on paired bootstrap resampling (Koehn, 2004) of best MERT runs for both configurations showed that Form→Form is significantly better (p-value=0.022) than Form+Sense→Form. The sentence-by-sentence comparison explains this: On the positive side, 8 examples out of the top 10 sentences where Form+Sense→Form output was better than Form→Form profited from using additional information about the verb sense. On the negative side, the model with verb senses made a lot of errors due to badly extracted phrase tables, even leaving some verbs untranslated.

3.1.2 Form→Form+Tag vs. Form+Sense→Form+Tag

In this case the same paired bootstrap resampling of the best MERT runs showed that the difference between Form+Sense→Form+Tag and Form→Form+Tag outputs is not significant (p-value=0.062). In the sentence by sentence comparison, we saw that while information about verb pattern helps to deal with some translations, it still causes mistakes.

For example, in the sentence from Figure 1, the verb *cool down* is translated as *vychladnout* ('let the temperature sink') instead of the correct *uklidnit* ('calm down'). Here, MT-ComparEval shows that Form→Form+Tag translated the verb correctly, meaning that the correct translation exists in the training data. Therefore, we checked which of the translation model factors caused the wrong translation. In the source sentence, the verb *cool* has the CPA pattern "1", but the only suitable phrase in the Form+Sense→Form+Tag phrase table (with `cool|1 down|-` on the source side) has the verb *vychladnout* on the target side. In the Form→Form+Tag table, we have the phrase `cool down and let` translated using the verb *uklidnit*, but the corresponding phrase in the Form+Sense→Form+Tag table has has a different CPA pattern "u" for the verb *cool*.

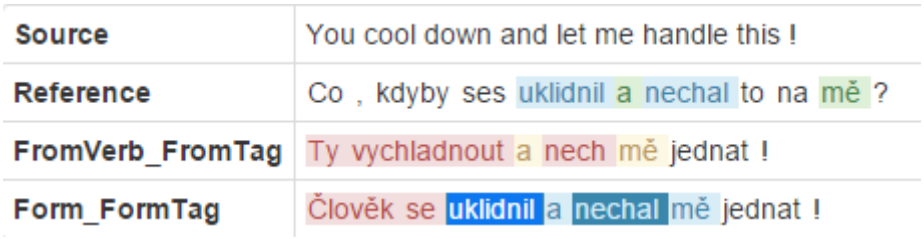

Figure 1: An example MT-ComparEval output from the Form+Sense→Form+Tag sentence analysis

work[1]: ACT PAT DIR3
(put, implement)
Burger King works a sales pitch into its public-service message.
work[2]: ACT ?PAT ?BEN ?ACMP
(perform a job)
Mr. Cray has been working on the project for more than six years.
work[3]: ACT PAT
(cause, create)
[. . .] greenhouse effect that will work important climatic changes [. . .]
work[4]: ACT
(function)
US trade law is working.

Figure 2: Example entry from the EngVallex valency dictionary, with four different senses/valency frames of the verb *work* (abridged, with minor adaptations for presentation).
The sense ID and the valency frame is shown on the 1st line of each sense, with the following semantic roles: ACT = actor, PAT = patient, DIR3 = direction (to, into), BEN = benefactor, ACMP = accompanying person or object. Optional arguments are prepended with a "?". A short gloss is shown on the 2nd line, and an example on the 3rd line.

3.1.3 Form→Form+Tag vs. Form→Form+Tag + Form+Sense→Form+Tag

The MT-ComparEval's paired bootstrap resampling showed that the difference between these two outputs is significant (p-value=0.023), thus showing that output of Form→Form+Tag + Form+Sense→Form+Tag is significantly better than Form→Form+Tag. In the sentence-by-sentence comparison, we saw that the combined system benefited from the verb patterns where possible but resorted to the more general translation of the baseline phrase-table when CPA-annotated translations were insufficient.

4 Verbal word senses in valency frames

Valency in verbs (and other parts of speech), i.e., the ability of a verb to require and shape its arguments, is one of the core notions of the Functional Generative Description (FGD) theory (Sgall et al., 1986). The valency of a verb is described in a valency frame, which lists the semantic roles and possible syntactic shapes of all of its obligatory and optional arguments. Since different senses of the same verb require different arguments and thus are described by different valency frames, this amounts to WSD in verbs (an example is shown in Figure 2).

Valency frames for over 7,000 senses of more than 4,000 common English verbs are listed in the EngVallex valency lexicon (Cinková, 2006),[3] and the Prague Czech-English Dependency Treebank (PCEDT) 2.0 (Hajič et al., 2012) provides manually annotated valency frame IDs for all of its verbs. Using this annotation, Dušek et al. (2015) trained an automatic system for valency frame detection as a part of the Treex natural language processing toolkit (Popel and Žabokrtský, 2010).[4] We processed all the sentences in our dataset with the tool and used the resulting valency frame IDs in our experiments.

4.1 Experiments and evaluation

Based on the results of the experiments shown in Section 3.1, we have decided to focus only on the following configurations: Form→Form+Tag, Form+Sense→Form+Tag and their combination

[3]EngVallex is origially based on the PropBank frame files (Palmer et al., 2005), but it also contains a lot of manual changes.
[4]`http://ufal.mff.cuni.cz/treex`

Configuration	BLEU
Form+Sense→Form+Tag	24.97
Form→Form+Tag	25.08
Form→Form+Tag + Form+Sense→Form+Tag	25.26

Table 5: Evaluation results for valency frames annotation, best MERT for each configuration

Form→Form+Tag + Form+Sense→Form+Tag.

Table 5 shows the results for best MERT runs for each configuration. MultEval MERT evaluation for the all configurations mentioned above, with Form→Form+Tag as a baseline, is shown in Table 6. The table shows that the average Form+Sense→Form+Tag model results are still 0.1% BLEU worse than the Form→Form+Tag model, but the average results of the combined Form→Form+Tag + Form+Sense→Form+Tag model are 0.1% BLEU better than the average results of Form→Form+Tag. The results of MultEval's stratified approximate randomization test (Clark et al., 2011) allow us to claim that the combination of these two models is statistically significantly better than the baseline. The same is true for METEOR and TER tests results, shown in the same table. It also shows that the valency frames approach to WSD has more impact on MT than CPA in our case.

Metric	**System**	**Avg**	$\bar{s}_{\text{sel}}$	s_{Test}	p**-value**
BLEU	Form→Form+Tag	25.0	0.9	0.1	-
	Form+Sense→Form+Tag	24.9	0.9	0.1	0.01
	Form→Form+Tag + Form+Sense→Form+Tag	25.1	0.9	0.1	0.00
METEOR	Form→Form+Tag	22.5	0.4	0.0	-
	Form+Sense→Form+Tag	22.5	0.4	0.0	0.01
	Form→Form+Tag + Form+Sense→Form+Tag	22.6	0.4	0.0	0.00
TER	Form→Form+Tag	62.2	0.7	0.1	-
	Form+Sense→Form+Tag	62.4	0.7	0.2	0.00
	Form→Form+Tag + Form+Sense→Form+Tag	62.1	0.7	0.2	0.00

Table 6: MultEval results for valency frames, based on 8 MERT runs

A more thorough examination of the best MERT runs of following pairs of configurations in MT-ComparEval output of paired bootstrap resampling showed that:

- Form+Sense→Form+Tag is insignificantly worse than Form→Form+Tag, with p-value=0.0161
- Form→Form+Tag + Form+Sense→Form+Tag is significantly better than Form→Form+Tag, with p-value=0.002

An interesting observation was that Form+Sense→Form+Tag and Form→Form+Tag + Form+Sense→Form+Tag models were more likely to translate verbs as verbs, while translation errors in Form→Form+Tag often were caused by its efforts to translate verbs as nouns.

4.2 Comparsion of CPA and valency frames

Based on the MultEval results shown in Table 4 and Table 6, it can be claimed that using the valency frames approach to WSD helped to achieve a statistically significant improvement in machine translation, while CPA did not help to such an extent. Among the possible reasons are a lower number of verbs covered (for the same number of sentences, we had CPA-based annotations only for 28 different verbs and 3,306 different verbs with valency frames annotations) and the precision of automatic annotating system itself. One of the future plans here is to compare the results of these approaches when exactly the same verbs are annotated.

An example of the sentence where the valency frames approach was more successful than CPA is "... forged steel components for the automotive industry". Here, the word *forged* was annotated by verbal valency frame and by verbal pattern, and while valency frame provided correct translation of this word into Czech "*kované* oceli součástí", the CPA-based model generated "*zfalšoval* ocel součástí", which is incorrect in both the meaning and the part of speech.

5 Discussion and conclusion

Including verb senses – be it based on corpus pattern analysis or as valency frames – as an additional factor to a PB-SMT English-to-Czech model did not help by itself, as our results for Form+Sense→Form+Tag configurations have shown. Nevertheless, the combination of this model with a better-performing model Form→Form+Tag resulted in a significant improvement for the case of using senses based on valency frames, as shown by paired boootstrap resampling tests given in Table 6, while a manual evaluation of best MERT runs showed translation quality improvement for both WSD approaches. All the results were achieved on a relatively small data sets, but it can be of use in cases when one does not have enough parallel data, but WSD for the source language (which is often English) is available, for example, in case of domain-specific translations.

We have tried to use sense information produced by two different approaches to verbal WSD disambiguation – corpus pattern analysis and valency frames, and while the former did not significantly outperform the baseline system in terms of the BLEU metric, the later showed significant improvement.

Adding the automatic WSD system as additional preprocessing layer can influence the SMT system due to the fact that WSD system cannot deliver 100% accurate senses, thus causing confusing situations, when the system had a correct translation available, but did not select it because the verb sense of the source sentence from test set was incorrect. Possible ways of reducing the impact of such things are improvement of automatic WSD systems used and using WSD system combination.

6 Future work

In the future, we plan to continue our experiments on verbs senses using approached described in this work as well as other approaches, e.g. WSD systems based on BabelNet synsets (Navigli and Ponzetto, 2012) and WordNet senses.[5] In addition, we are going to experiment with the size of the corpus used for training, because this research used only a part of available Czech-English parallel corpus.

7 Acknowledgments

This research was supported by the grants H2020-ICT-2014-1-645452, GBP103/12/G084, SVV 260 333, and using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071). We thank the two anonymous reviewers for useful comments.

References

Marianna Apidianaki, Benjamin Marie, and Lingua et Machina. 2015. METEOR-WSD: improved sense matching in MT evaluation. *Syntax, Semantics and Structure in Statistical Translation*, page 49.

Nora Aranberri, Eleftherios Avramidis, Aljoscha Burchardt, Ondrej Klejch, Martin Popel, and Maja Popovic. 2016. Tools and guidelines for principled machine translation development. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1877–1882, Portorož, Slovenia.

BNC. 2007. British national corpus, version 3 (BNC XML edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Ondřej Bojar and Aleš Tamchyna. 2013. The Design of Eman, an Experiment Manager. *The Prague Bulletin of Mathematical Linguistics*, 99:39–58.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proceedings of LREC 2012*, Istanbul, Turkey, May. ELRA, European Language Resources Association. In print.

Ondřej Bojar. 2007. English-to-Czech Factored Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 232–239, Prague, Czech Republic, June. Association for Computational Linguistics.

[5] `http://wordnet.princeton.edu`

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Prague, Czech Republic.

Silvie Cinková, Martin Holub, Adam Rambousek, and Lenka Smejkalová. 2012. A database of semantic clusters of verb usages. In *Proceedings of the LREC 2012 International Conference on Language Resources and Evaluation*. Istanbul, Turkey.

Silvie Cinková. 2006. From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the fifth International conference on Language Resources and Evaluation (LREC 2006), Genova, Italy*.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181. Association for Computational Linguistics.

Ondřej Dušek, Eva Fučíková, Jan Hajič, Martin Popel, Jana Šindlerová, and Zdeňka Urešová. 2015. Using Parallel Texts and Lexicons for Verbal Word Sense Disambiguation. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 82–90, Uppsala, Sweden.

J. Hajič, E. Hajičová, J. Panevová, P. Sgall, O. Bojar, S. Cinková, E. Fučíková, M. Mikulová, P. Pajas, J. Popelka, J. Semecký, J. Šindlerová, J. Štěpánek, J. Toman, Z. Urešová, and Z. Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of LREC*, pages 3153–3160, Istanbul.

Patrick Hanks and James Pustejovsky. 2005. A Pattern Dictionary for Natural Language Processing. *Revue Francaise de linguistique appliquée*, 10(2).

Patrick Hanks. 1994. Linguistic norms and pragmatic exploitations, or why lexicographers need prototype theory and vice versa. In F. Kiefer, G. Kiss, and J. Pajzs, editors, *Papers in Computational Lexicography: Complex '94*. Research Institute for Linguistics, Hungarian Academy of Sciences.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. University Press Group Limited.

Alice F Healy and George A Miller. 1970. Verb as main determinant of sentence meaning. *Psychonomic Science*, 20(6):372–372.

Martin Holub, Vincent Kríz, Silvie Cinková, and Eckhard Bick. 2012. Tailored feature extraction for lexical disambiguation of english verbs based on corpus pattern analysis. In *COLING*, pages 1195–1210.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proc. of EMNLP*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, pages 187–193.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395. Citeseer.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Steven Neale, Luıs Gomes, and António Branco. 2015. First steps in using word senses as contextual features in maxent models for machine translation. In *1st Deep Machine Translation Workshop*, page 64.

Steven Neale, Luıs Gomes, Eneko Agirre, Oier Lopez de Lacalle, and António Branco. 2016. Word sense-aware machine translation: Including senses as contextual features for improved translation models. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2777–2783, Portorož, Slovenia.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: modular NLP framework. In *Proceedings of IceTAL, 7th International Conference on Natural Language Processing*, pages 293–304, Reykjavík.

P. Sgall, E. Hajičová, and J. Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. D. Reidel, Dordrecht.