Machine Translation of Medical Texts in the Khresmoi Project

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Michal Novák,
Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, Daniel Zeman
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 11800 Prague, Czech Republic
{odusek,hajic,hlavacova,mnovak,pecina,rosa,tamchyna,uresova,zeman}@ufal.mff.cuni.cz

Abstract

This paper presents the participation of the Charles University team in the WMT 2014 Medical Translation Task. Our systems are developed within the Khresmoi project, a large integrated project aiming to deliver a multi-lingual multi-modal search and access system for biomedical information and documents. Being involved in the organization of the Medical Translation Task, our primary goal is to set up a baseline for both its subtasks (summary translation and query translation) and for all translation directions. Our systems are based on the phrase-based Moses system and standard methods for domain adaptation. The constrained/unconstrained systems differ in the training data only.

1 Introduction

The WMT 2014 Medical Translation Task poses an interesting challenge for Machine Translation (MT). In the "standard" translation task, the end application is the translation itself. In the Medical Translation Task, the MT system is considered a part of a larger system for Cross-Lingual Information Retrieval (CLIR) and is used to solve two different problems: (i) translation of user search queries, and (ii) translation of summaries of retrieved documents.

In query translation, the end user does not even necessarily see the MT output as their queries are translated and search is performed on documents in the target language. In summary translation, the sentences to be translated come from document summaries (snippets) displayed to provide information on each of the documents retrieved by the search. Therefore, translation quality may not be the most important measure in this task – the performance of the CLIR system as a whole is the final criterion. Another fundamental difference from the standard task is the nature of the translated texts. While we can consider document summaries to be ordinary texts (despite their higher information density in terms of terminology from a narrow domain), search queries in the medical domain are an extremely specific type of data, and traditional techniques for system development and domain adaptation are truly put to a test here.

This work is a part of the of the large integrated EU-funded Khresmoi project.[1] Among other goals, such as joint text and image retrieval of radiodiagnostic records, Khresmoi aims to develop technology for transparent cross-lingual search of medical sources for both professionals and laypeople, with the emphasis primarily on publicly available web sources.

In this paper, we describe the Khresmoi systems submitted to the WMT 2014 Medical Translation Task. We participate in both subtasks (summary translation and query translation) for all language pairs (Czech–English, German–English, and French–English) in both directions (to English and from English). Our systems are based on the Moses phrase-based translation toolkit and standard methods for domain adaptation. We submit one constrained and one unconstrained system for each subtask and translation direction. The constrained and unconstrained systems differ in training data only: The former use all allowed training data, the latter take advantage of additional web-crawled data.

We first summarize previous works in MT domain adaptation in Section 2, then describe the data we used for our systems in Section 3. Sec-

[1] http://www.khresmoi.eu/

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 221–228,
Baltimore, Maryland USA, June 26–27, 2014. ©2014 Association for Computational Linguistics

tion 4 contains an account of the submitted systems and their performance in translation of search queries and document summaries. Section 5 concludes the paper.

2 Related work

To put our work in the context of other approaches, we first describe previous work on domain adaptation in Statistical Machine Translation (SMT), then focus specifically on SMT in the medical domain.

2.1 Domain adaptation of Statistical machine translation

Many works on domain adaptation examine the usage of available in-domain data to directly improve in-domain performance of SMT. Some authors attempt to combine the predictions of two separate (in-domain and general-domain) translation models (Langlais, 2002; Sanchis-Trilles and Casacuberta, 2010; Bisazza et al., 2011; Nakov, 2008) or language models (Koehn and Schroeder, 2007). Wu and Wang (2004) use in-domain data to improve word alignment in the training phase. Carpuat et al. (2012) explore the possibility of using word sense disambiguation to discriminate between domains.

Other approaches concentrate on the acquisition of larger in-domain corpora. Some of them exploit existing general-domain corpora by selecting data that resemble the properties of in-domain data (e.g., using cross-entropy), thus building a larger *pseudo-in-domain* training corpus. This technique is used to adapt language models (Eck et al., 2004b; Moore and Lewis, 2010) as well as translation models (Hildebrand et al., 2005; Axelrod et al., 2011) or their combination (Mansour et al., 2011). Similar approaches to domain adaptation are also applied in other tasks, e.g., automatic speech recognition (Byrne et al., 2004).

2.2 Statistical machine translation in the medical domain

Eck et al. (2004a) employ an SMT system for the translation of dialogues between doctors and patients and show that according to automatic metrics, a dictionary extracted from the Unified Medical Language System (UMLS) Metathesaurus and its semantic type classification (U.S. National Library of Medicine, 2009) significantly improves translation quality from Spanish to English when applied to generalize the training data.

Wu et al. (2011) analyze the quality of MT on PubMed[2] titles and whether it is sufficient for patients. The conclusions are very positive especially for languages with large training resources (English, Spanish, German) – the average fluency and content scores (based on human evaluation) are above four on a five-point scale. In automatic evaluation, their systems substantially outperform Google Translate. However, the SMT systems are specifically trained, tuned, and tested on the domain of PubMed titles, and it is not evident how they would perform on other medical texts.

Costa-jussà et al. (2012) are less optimistic regarding SMT quality in the medical domain. They analyze and evaluate the quality of public web-based MT systems (such as Google Translate) and conclude that in both automatic and manual evaluation (on 7 language pairs), the performance of these systems is still not good enough to be used in daily routines of medical doctors in hospitals.

Jimeno Yepes et al. (2013) propose a method for obtaining in-domain parallel corpora from titles and abstracts of publications in the MEDLINE[3] database. The acquired corpora contain from 30,000 to 130,000 sentence pairs (depending on the language pair) and are reported to improve translation quality when used for SMT training, compared to a baseline trained on out-of-domain data. However, the authors use only one source of in-domain parallel data to adapt the translation model, and do not use any in-domain monolingual data to adapt the language model.

In this work, we investigate methods combining the different kinds of data – general-domain, in-domain, and pseudo-in-domain – to find the optimal approach to this problem.

3 Data description

This section includes an overview of the parallel and monolingual data sources used to train our systems. Following the task specification, they are split into constrained and unconstrained sections. The constrained section includes medical-domain data provided for this task (extracted by the provided scripts), and general-domain texts provided as constrained data for the standard task ("general domain" here is used to denote data

[2] http://www.ncbi.nlm.nih.gov/pubmed/

[3] http://www.nlm.nih.gov/pubs/factsheets/medline.html

		Czech–English			German–English			French–English		
dom	set	pairs	source	target	pairs	source	target	pairs	source	target
med	con	2,498	18,126	19,964	4,998	123,686	130,598	6,139	202,245	171,928
gen	con	15,788	226,711	260,505	4,520	112,818	119,404	40,842	1,470,016	1,211,516
gen	unc	–	–	–	9,320	525,782	574,373	13,809	961,991	808,222

Table 1: Number of sentence pairs and tokens (source/target) in parallel training data (in thousands).

dom	set	English	Czech	German	French
med	con	172,991	1,848	63,499	63,022
gen	con	6,132,107	627,493	1,728,065	1,837,457
med	unc	3,275,272	36,348	361,881	908,911
gen	unc	618,084	–	339,595	204,025

Table 2: Number of tokens in monolingual training data (in thousands).

which comes from a mixture of various different domains, mostly news, parliament proceedings, web-crawls, etc.). The unconstrained section contains automatically crawled data from medical and health websites and non-medical data from patent collections.

3.1 Parallel data

The parallel data summary is presented in Table 1.

The main sources of the medical-domain data for all the language pairs include the EMEA corpus (Tiedemann, 2009), the UMLS metathesaurus of health and biomedical vocabularies and standards (U.S. National Library of Medicine, 2009), and bilingual titles of Wikipedia articles belonging to the categories identified to be medical domain. Additional medical-domain data comes from the MAREC patent collection: PatTR (Wäschle and Riezler, 2012) available for DE–EN and FR–EN, and COPPA (Pouliquen and Mazenc, 2011) for FR–EN (only patents from the medical categories A61, C12N, and C12P are allowed in the constrained systems).

The constrained general-domain data include three parallel corpora for all the language pairs: CommonCrawl (Smith et al., 2013), Europarl version 6 (Koehn, 2005), the News Commentary corpus (Callison-Burch et al., 2012). Further, the constrained data include CzEng (Bojar et al., 2012) for CS–EN and the UN corpus for FR–EN.

For our unconstrained experiments, we also employ parallel data from the non-medical patents from the PatTR and COPPA collections (other categories than A61, C12N, and C12P).

3.2 Monolingual data

The monolingual data is summarized in Table 2.

The main sources of the medical-domain monolingual data for all languages involve Wikipedia pages, UMLS concept descriptions, and non-parallel texts extracted from the medical patents of the PatTR collections. For English, the main source is the AACT collection of texts from ClinicalTrials.gov. Smaller resources include: DrugBank (Knox et al., 2011), GENIA (Kim et al., 2003), FMA (Rosse and Mejino Jr., 2008), GREC (Thompson et al., 2009), and PIL (Bouayad-Agha et al., 2000).

In the unconstrained systems, we use additional monolingual data from web pages crawled within the Khresmoi project: a collection of about one million HON-certified[4] webpages in English released as the test collection for the CLEF 2013 eHealth Task 3 evaluation campaign,[5] additional web-crawled HON-certified pages (not publicly available), and other webcrawled medical-domain related webpages.

The constrained general-domain resources include: the News corpus for CS, DE, EN, and FR collected for the purpose of the WMT 2014 Standard Task, monolingual parts of the Europarl and News-Commentary corpora, and the Gigaword for EN and FR.

For the FR–EN and DE–EN unconstrained systems, the additional general domain monolingual data is taken from monolingual texts of non-medical patents in the PatTR collection.

[4] https://www.hon.ch/

[5] https://sites.google.com/site/shareclefehealth/

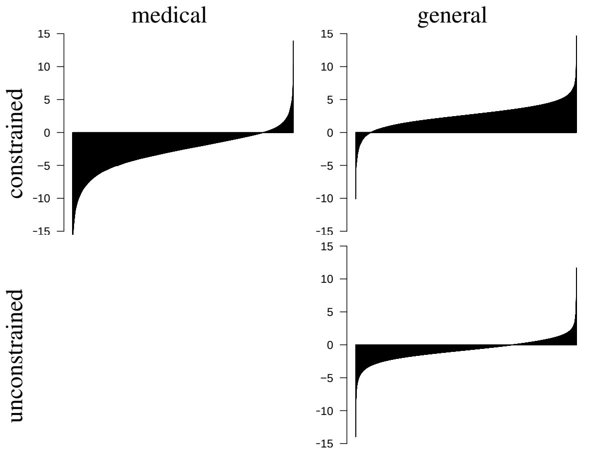

Figure 1: Distribution of the domain-specificity scores in the English–French parallel data sets.

3.3 Data preprocessing

The data consisting of crawled web pages, namely CLEF, HON, and non-HON, needed to be cleaned and transformed into a set of sentences. The Boilerpipe (Kohlschütter et al., 2010) and Justext (Pomikálek, 2011) tools were used to remove boilerplate texts and extract just the main content from the web pages. The YALI language detection tool (Majliš, 2012) trained on both in-domain and general domain data then filtered out those cleaned pages which were not identified as written in one of the concerned languages.

The rest of the preprocessing procedure was applied to all the datasets mentioned above, both parallel and monolingual. The data were tokenized and normalized by converting or omitting some (mostly punctuation) characters. A set of language-dependent heuristics was applied in an attempt to restore and normalize the opening/closing quotation marks, i.e. convert "*quoted*" to “*quoted*” (Zeman, 2012). The motivation here is twofold: First, we hope that paired quotation marks could occasionally work as brackets and better denote parallel phrases for Moses; second, if Moses learns to output directed quotation marks, the subsequent detokenization will be easier. For all systems which translate *from* German, decompounding is employed to reduce source-side data sparsity. We used BananaSplit for this task (Müller and Gurevych, 2006).

We perform all training and internal evaluation on lowercased data; we trained recasers to post-process the final submissions.

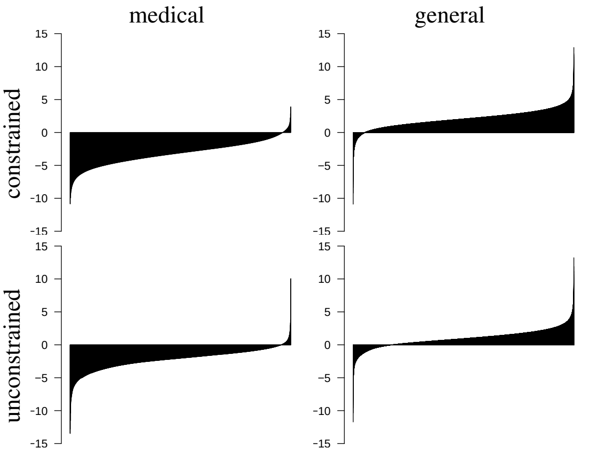

Figure 2: Distribution of the domain-specificity scores in the French monolingual data sets.

4 Submitted systems

We first describe our technique of psedo-in-domain data selection in Section 4.1, then compare two methods of combining the selected data in Section 4.2. This, along with using constrained and unconstrained data sets to train the systems (see Section 3), amounts to a total of four system variants submitted for each task. A description of the system settings used is given in Section 4.3.

4.1 Data selection

We follow an approach originally proposed for selection of monolingual sentences for language modeling (Moore and Lewis, 2010) and its modification applied to selection of parallel sentences (Axelrod et al., 2011). This technique assumes two language models for sentence scoring, one trained on (true) in-domain text and one trained on (any) general-domain text in the same language (e.g., English). For both data domains (general and medical), we score each sentence by the difference of its cross-perplexity given the in-domain language model and cross-perplexity given the general-domain language model (in this order). We only keep sentences with a *negative score* in our data, assuming that these are the most “medical-like”. Visualisation of the domain-specificity scores (cross-perplexity difference) in the FR–EN parallel data and FR monolingual data is illustrated in Figures 1 and 2, respectively.[6] The scores (Y axis) are presented for each sentence in increasing order from left to right (X axis).

[6] For the medical domain, constrained and unconstrained parallel data are identical.

		cs→en	de→en	en→cs	en→de	en→fr	fr→en
con	concat	33.64±1.14	32.84±1.24	18.10±0.94	18.29±0.92	33.39±1.11	36.71±1.17
con	interpol	32.94±1.11	32.31±1.20	18.96±0.93	18.41±0.93	34.06±1.11	37.42±1.21
unc	concat	34.10±1.11	34.52±1.20	21.12±1.03	19.76±0.92	36.23±1.03	**38.15±1.16**
unc	interpol	**34.48±1.16**	**34.92±1.17**	**22.15±1.06**	**20.81±0.95**	**36.26±1.13**	37.91±1.13

Table 3: BLEU scores of summary translations.

		cs→en	de→en	en→cs	en→de	en→fr	fr→en
con	concat	30.87±4.70	33.21±5.03	23.25±4.85	17.72±4.75	28.64±3.77	35.56±4.94
con	interpol	32.46±5.05	33.74±4.97	21.56±4.80	16.90±4.39	29.34±3.73	35.28±5.26
unc	concat	**34.88±5.04**	31.24±5.59	22.61±4.91	**19.13±5.66**	**33.08±3.80**	36.73±4.88
unc	interpol	33.82±5.16	**34.19±5.27**	**23.93±5.16**	15.87±11.31	31.19±3.73	**40.25±5.14**

Table 4: BLEU scores of query translations.

The two language models for sentence scoring are trained with a restricted vocabulary extracted from the in-domain training data as words occurring at least twice (singletons and other words are treated as out-of-vocabulary). In our experiments, we apply this technique to select both monolingual data for language models and parallel data for translation models. Selection of parallel data is based on the English side only. The in-domain models are trained on the monolingual data in the target language (constrained or unconstrained, depending on the setting). The general-domain models are trained on the WMT News data.

Compared to the approach of Moore and Lewis (2010) and Axelrod et al. (2011), we prune the model vocabulary more aggressively – we discard not only the singletons, but also all words with non-Latin characters, which helps clean the models from noise introduced by the automatic process of data acquisition by web crawling.

4.2 Data combination

For both parallel and monolingual data, we obtain two data sets after applying the data selection:

- “medical-like” data from the medical domain
- “medical-like” data from the general domain.

For each language pair and for each system type (constrained/unconstrained), we submitted two system variants which differ in how the selected data are combined. The first variant uses a simple concatenation of the two datasets both for parallel data and for language model data. In the second variant, we train separate models for each section and use *linear interpolation* to combine them into a single model. For language models, we use the SRILM linear interpolation feature (Stolcke, 2002). We interpolate phrase tables using Tmcombine (Sennrich, 2012). In both cases, the held-out set for minimizing the perplexity is the system development set.

4.3 System details

We compute word alignment on lowercase 4-character stems using fast_align (Dyer et al., 2013). We create phrase tables using the Moses toolkit (Koehn et al., 2007) with standard settings. We train 5-gram language models on the target-side lowercase forms using SRILM. We use MERT (Och, 2003) to tune model weights in our systems on the development data provided for the task.

The only difference between the system variants for query and summary translation is the tuning set. In both cases, we use the respective sets provided offcially for the shared task.

4.4 Results

Tables 3 and 4 show case-insensitive BLEU scores of our systems.[7] As expected, the unconstrained systems outperform the constrained ones. Linear interpolation outperforms data concatenation quite reliably across language pairs for summary translation. While the picture for query translation is similar, there is more variance in the results, so we cannot state that interpolation definitely works

[7] As we use the same recasers for both summary and query translation, our systems are heavily penalized for wrong letter case in query translation. However, letter case is not taken into account in most CLIR systems. All BLEU scores reported in this paper will be case-insensitive for this reason.

better in this case. This is due to the sizes of the development and test sets and most importantly due to sentence lengths – queries are very short, making BLEU unreliable, MERT unstable, and bootstrap resampling intervals wide.

If we compare our score to the other competitors, we are clearly worse than the best systems for summary translation. From this perspective, our data filtering seems overly eager (i.e., discarding all sentence pairs with a positive perplexity difference). An experiment which we leave for future work is doing one more round of interpolation to combine a model trained on the data with negative perplexity with models trained on the remainder.

5 Conclusions

We described the Charles University MT system used in the Shared Medical Translation Task of WMT 2014. Our primary goal was to set up a baseline for both the subtasks and all translation directions. The systems are based on the Moses toolkit, pseudo-in-domain data selection based on perplexity difference and two different methods of in-domain and out-of-domain data combination: simple data concatenation and linear model interpolation.

We report results of constrained and unconstrained systems which differ in the training data only. In most experiments, using additional data improved the results compared to the constrained systems and using linear model interpolation outperformed data concatenation. While our systems are on par with best results for case-insensitive BLEU score in query translation, our overly eager data selection techniques caused lower scores in summary translation. In future work, we plan to include a special out-of-domain model in our setup to compensate for this problem.

Acknowledgments

This work was supported by the EU FP7 project Khresmoi (contract no. 257528), the Czech Science Foundation (grant no. P103/12/G084), and SVV project number 260 104. This work has been using language resources developed, stored, and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

References

A. Axelrod, X. He, and J. Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, United Kingdom. ACL.

A. Bisazza, N. Ruiz, and M. Federico. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 136–143, San Francisco, CA, USA. International Speech Communication Association.

O. Bojar, Z. Žabokrtský, O. Dušek, P. Galuščáková, M. Majliš, D. Mareček, J. Maršík, M. Novák, M. Popel, and A. Tamchyna. 2012. The joy of parallelism with CzEng 1.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 3921–3928, Istanbul, Turkey. European Language Resources Association.

N. Bouayad-Agha, D. R. Scott, and R. Power. 2000. Integrating content and style in documents: A case study of patient information leaflets. *Information Design Journal*, 9(2–3):161–176.

W. Byrne, D. S. Doermann, M. Franz, S. Gustman, J. Hajič, D. W. Oard, et al. 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *Speech and Audio Processing, IEEE Transactions on*, 12(4):420–435.

C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. ACL.

M. Carpuat, H. Daumé III, A. Fraser, C. Quirk, F. Braune, A. Clifton, et al. 2012. Domain adaptation in machine translation: Final report. In *2012 Johns Hopkins Summer Workshop Final Report*, pages 61–72. Johns Hopkins University.

M. R. Costa-jussà, M. Farrús, and J. Serrano Pons. 2012. Machine translation in medicine. A quality analysis of statistical machine translation in the medical domain. In *Proceedings of the 1st Virtual International Conference on Advanced Research in Scientific Areas*, pages 1995–1998, Žilina, Slovakia. Žilinská univerzita.

C. Dyer, V. Chahuneau, and N. A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of NAACL-HLT*, pages 644–648.

M. Eck, S. Vogel, and A. Waibel. 2004a. Improving statistical machine translation in the medical domain using the Unified Medical Language System. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 792–798, Geneva, Switzerland. ACL.

M. Eck, S. Vogel, and A. Waibel. 2004b. Language model adaptation for statistical machine translation based on information retrieval. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the International Conference on Language Resources and Evaluation*, pages 327–330, Lisbon, Portugal. European Language Resources Association.

A. S. Hildebrand, M. Eck, S. Vogel, and A. Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, pages 133–142, Budapest, Hungary. European Association for Machine Translation.

A. Jimeno Yepes, É. Prieur-Gaston, and A. Névéol. 2013. Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC Bioinformatics*, 14(1):1–10.

J.-D Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

C. Knox, V. Law, T. Jewison, P. Liu, Son Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart. 2011. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic acids research*, 39(suppl 1):D1035–D1041.

P. Koehn and J. Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic. ACL.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Praha, Czechia, June. ACL.

P. Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. Asia-Pacific Association for Machine Translation.

C. Kohlschütter, P. Fankhauser, and W. Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 441–450, New York, NY, USA. ACM.

P. Langlais. 2002. Improving a general-purpose statistical translation engine by terminological lexicons. In *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology*, volume 14, pages 1–7, Taipei, Taiwan. ACL.

M. Majliš. 2012. Yet another language identifier. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 46–54, Avignon, France. ACL.

S. Mansour, J. Wuebker, and H. Ney. 2011. Combining translation and language model scoring for domain-specific data filtering. In *International Workshop on Spoken Language Translation*, pages 222–229, San Francisco, CA, USA. ISCA.

R. C. Moore and W. Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. ACL.

C. Müller and I. Gurevych. 2006. Exploring the potential of semantic relatedness in information retrieval. In *LWA 2006 Lernen – Wissensentdeckung – Adaptivität, 9.-11.10.2006, Hildesheimer Informatikberichte*, pages 126–131, Hildesheim, Germany. Universität Hildesheim.

P. Nakov. 2008. Improving English–Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 147–150, Columbus, OH, USA. ACL.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. ACL.

J. Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD thesis, Masaryk University, Faculty of Informatics, Brno.

B. Pouliquen and C. Mazenc. 2011. COPPA, CLIR and TAPTA: three tools to assist in overcoming the patent barrier at WIPO. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 24–30, Xiamen, China. Asia-Pacific Association for Machine Translation.

C. Rosse and José L. V. Mejino Jr. 2008. The foundational model of anatomy ontology. In A. Burger, D. Davidson, and R. Baldock, editors, *Anatomy Ontologies for Bioinformatics*, volume 6 of *Computational Biology*, pages 59–117. Springer London.

G. Sanchis-Trilles and F. Casacuberta. 2010. Log-linear weight optimisation via Bayesian adaptation in statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1077–1085, Beijing, China. ACL.

R. Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549. ACL.

J. R. Smith, H. Saint-Amand, M. Plamada, P. Koehn, C. Callison-Burch, and A. Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. ACL.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, Denver, Colorado, USA.

P. Thompson, S. Iqbal, J. McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC bioinformatics*, 10(1):349.

J. Tiedemann. 2009. News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248, Borovets, Bulgaria. John Benjamins.

U.S. National Library of Medicine. 2009. UMLS reference manual. Metathesaurus. Bethesda, MD, USA.

K. Wäschle and S. Riezler. 2012. Analyzing parallelism and domain similarities in the MAREC patent corpus. In M. Salampasis and B. Larsen, editors, *Multidisciplinary Information Retrieval*, volume 7356 of *Lecture Notes in Computer Science*, pages 12–27. Springer Berlin Heidelberg.

H. Wu and H. Wang. 2004. Improving domain-specific word alignment with a general bilingual corpus. In Robert E. Frederking and Kathryn B. Taylor, editors, *Machine Translation: From Real Users to Research*, volume 3265 of *Lecture Notes in Computer Science*, pages 262–271. Springer Berlin Heidelberg.

C. Wu, F. Xia, L. Deleger, and I. Solti. 2011. Statistical machine translation for biomedical text: are we there yet? *AMIA Annual Symposium proceedings*, pages 1290–1299.

D. Zeman. 2012. Data issues of the multilingual translation matrix. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 395–400, Montréal, Canada. ACL.