Text Genre – An Unexplored Parameter in Statistical Machine Translation

Monica Gavrila and Cristina Vertan(✉)

University of Hamburg, 30 Vogt-Koelln Str, 22527 Hamburg, Germany
gavrila@informatik.uni-hamburg.de,
cristina.vertan@uni-hamburg.de
http://www.informatik.uni-hamburg.de

Abstract. It is generally accepted that the performance of a statistical machine translation (SMT) system depends significantly on the concordance between the domain of training and test data. During the last years several methods have been proposed in order to deal with out- of-domain words. Less to no attention has been paid however to text genre within the same domain. In this paper we demonstrate that the style of the training corpus may influence the quality of the translation output even when the domain of the training and test data remains al- most unchanged, but the text genre changes. We use as training data the JRC-Acquis and as test data the Europarl corpus. We include also experiments with an out-of-domain test data, as comparison for the variation of performance of the SMT system.

Keywords: Statistical machine translation · Text genre · Europarl · JRC-Acquis · RoGER · SMT evaluation

1 Introduction

It is generally accepted that statistical machine translation (SMT) provides sufficiently good translation results with in-domain test data and "enough" training data. Results are rapidly decreasing for out-of-domain test data. Therefore, lot of research has been directed in the last years towards domain adaptation of SMT systems - e.g. [10]. Especially for European languages, current state-of-the-art SMT-engines are trained on one of the two large corpora available: JRC-Acquis[1] or Europarl[2]. Special techniques are applied in a second phase in order to ensure lexical domain adaptation. Less attention is paid to the fact that, even inside one domain, corpora belong to different text genres or, at least, have different discourse structures and, therefore, other types of syntactic structures or semantic frames. These differences may have a bigger influence on the quality of an SMT-system than assumed until now.

1.1 The Context for the Experiments

This aspect is of particular importance in scenarios where a machine translation engine is part of a complex architecture exposed to textual input from heterogeneous

[1] http://ipsc.jrc.ec.europa.eu/index.php?id=198
[2] http://www.statmt.org/europarl/

Z. Vetulani and J. Mariani (Eds.): LTC 2011, LNAI 8387, pp. 456–467, 2014.
DOI: 10.1007/978-3-319-08958-4_37

domains or text genres. This is the case of a Web- Content-Management System (WCMS) as the ATLAS System[3].

In this system several web-services based on advanced language technology components are built for seven European languages. Among the key technologies which are incorporated, a central role is played by machine translation. Due to lack of enough training data for all possible domains, the data-driven translation engine is trained mostly on the JRC-Acquis corpus and afterwards domain adaptation is performed. For domains for which no training model is available, the user is informed that the translation quality can lack accuracy.

As the acceptance of such system depends extensively from the user acceptance we decided to investigate also to which extent the text genre of the input can influence the translation quality.

This paper shows several SMT experiments with different test data (in- domain vs. out-of-domain vs. 'similar' data) using the JRC-Acquis corpus for training. The language-pair considered is English-Romanian. The originality of the work is not in the MT approach involved, but in the way of choosing the test data. SMT experiments using JRC-Acquis and Romanian-English as language pair have been presented in [2, 6] and [9]. The results are presented in a tabular form in Table 1.

Table 1. Previous reported results

Direction of translation	Paper	BLEU score
English-Romanian	[2]	0.5464
	[6]	0.3208
	[9]	0.4900
Romanian-English	[2]	0.4604
	[6]	0.3840
	[9]	0.6080

SMT experiments have been usually performed and presented with in-domain data, for example see the experiments from [9] or [6].

An overview of how (rule-based) machine translation (MT) reacts to various text genres is shown in [1], where the MT system used is SYSTRAN[4]. The study analyzed machine translated extracts from four text genres with respect to different linguistic errors. Best results were obtained for technical sets of instructions.

Our paper is organized in seven sections. After this short introduction we will present the environmentof the MT-Engine in Sect. 2, while in Sect. 3 we describe our experimental settings: the MT system and the training and test data.

In Sect. 4 we show the evaluation results, followed in Sect. 5 by presenting factors which influence the results. The paper presents the conclusions and further work in Sect. 6. The last part of the paper shows our acknowledgments.

[3] http://www.atlasproject.eu

[4] http://www.systranet.com/

2 The ATLAS Content Management System

The core online service of the ATLAS platform is i-Publisher, a powerful Web-based instrument for creating, running and managing content-driven Web sites. It integrates the language-based technology to improve content navigation e.g. by interlinking documents based on extracted phrases, words and names, providing short summaries and suggested categorization concepts. Currently two different thematic content-driven Web sites, i-Librarian and EUDocLib, are being built on top of ATLAS platform, using i-Publisher as content management layer. i-Librarian is intended to be a user-oriented web site which allows visitors to maintain a personal workspace for storing, sharing and publishing various types of documents and have them automatically categorized into appropriate subject categories, summarized and annotated with important words, phrases and names. EUDocLib is planned as a publicly accessible repository of EU legal documents from the EUR-LEX collection with enhanced navigation and multilingual access.

The i-Publisher service:

- is mainly targeted at small enterprises and non-profit organizations,
- gives the ability to build via point-and-click user interface content-driven Web sites, which provide a wide set of pre-defined functionalities and the textual content of which is automatically processed, i.e. categorized, summarized, annotated, etc.,
- enables publishers, information designers and graphic designers to easily collaborate,
- aims at saving authors, editors and other contributors valuable time by automatically processing textual data and allows them to work together to produce high quality content. The last evaluation round of the service indicates that users do really see the benefit of LT-Technologies embedded into the system

The i-Librarian service:

- addresses the needs of authors, students, young researchers and readers,
- gives the ability to easily create, organize and publish various types of documents,
- allows users to find similar documents in different languages, to share personal works with other people, and to locate the most relevant texts from large collections of unfamiliar documents.

The EUDocLib service is a particular refinement of i-Librarian targeted to the management of documents from the European Commission.

The facilities described above are supported through intelligent language technology components like automatic classification, named entity recognition and information extraction, automatic text summarization, machine translation and cross-lingual retrieval. These components are integrated into the system in brick-like architecture, which means that each component is building on top of the other. The baseline brick is the language processing chains component which ensure a heterogonous linguistic processing of all documents independent of their language. A processing chain for a given language includes a number of existing tools, adjusted and/or fine-tuned to ensure their interoperability. In most respects a language

processing chain does not require development of new software modules but rather combining existing tools. The basic ATLAS software[5] is distributed as a software package under GPL license. LT-plug-ins like e.g. the language processing chains or the MT-engine follow a commercial licensing. The iLibrarian is available as web-service and it has unrestricted access.

2.1 Machine Translation in ATLAS System

Machine Translation is a key component of the ATLAS system. The development of the engine is particular challenging as the translation should be used in different domains. Additionally, the considered language-pairs belong to less resourced group[6], for which bilingual training and test material is available in limited amount.

The machine translation engine is integrated in 2 distinct ways into the ATLAS platform:

- for i-Publisher Service (generic platform for generating websites) the MT is serving as a translation aid tool for publishing multilingual content. Text is submitted to the translation engine and the result is subject to the human post processing
- for i-Librarian and EuDocLib (dedicated web services for collecting documents) the MT-engine provides a translation for assimilation, which means that the user retrieving documents in different languages will use the engine in order to get a clue about the documents, and decide if he will store them. If the translation is considered as acceptable it will be stored into a database.

The integration of a machine translation engine into a web based content management system in general and the ATLAS system in particular, presents from the user point of view several challenges among which we mention two, which ATLAS-System dealt with

1. The user may retrieve documents from different domains. Domain adaptation is a major issue in machine translation, and in particular in corpus–based methods. Poor lexical coverage and false disambiguation are the main issues when translating documents out of the training domain
2. The user may retrieve documents from various time periods. As language changes over time, language technology tools developed for the modern languages do not work equally well on diachronic documents.

With the current available technology it is not possible to provide a translation system which is domain and language variation independent and works for a couple of heterogeneous language pairs. Therefore our approach envisages a system of user guidance, so that the availability and the foreseen system-performance are transparent at any time.

[5] http://atlasproject.eu

[6] see http://www.meta-net.eu/whitepapers

For the MT-Engine of the ATLAS system we decided on a hybrid architecture combining EBMT [4] and SMT [8] at word-based level (no syntactic trees will be used). An original approach of our system is the interaction of the MT-engine with other modules of the system:

- The document categorization module assigns to each document one or more domains. For each domain the system administrator has the possibility to store information regarding the availability of a correspondent specific training corpus. If no specific trained model for the respective domain exists, the user is provided with a warning, telling that the translation may be inadequate with respect to the lexical coverage.
- The output of the summarization module is processed in such a way that ellipses and anaphora are omitted, and lexical material is adapted to the training corpus.

The information extraction module is providing information about metadata of the document including publication age. For documents previous to 1900 we will not provide translation, explaining the user that in absence of a training corpus the translation may be misleading.

The domain and dating restrictions can be changed at any time by the system administrator when an adequate training model is provided. The described architecture is presented in Fig. 1.

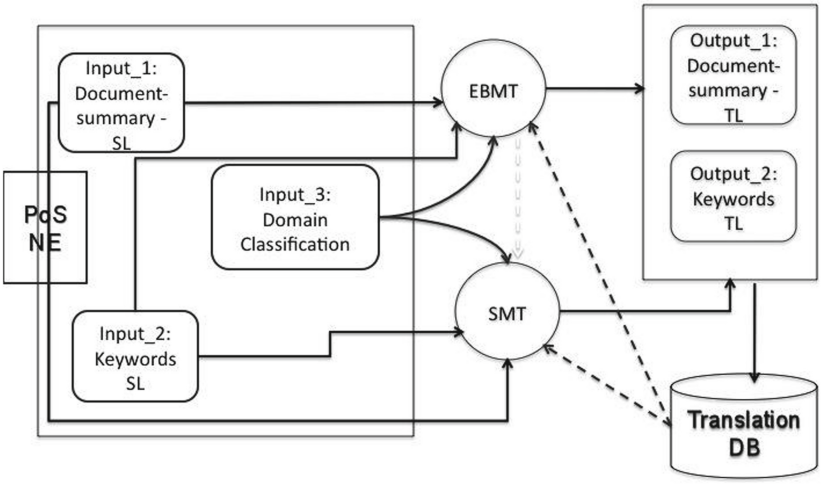

Fig. 1. System architecture for ATLAS-Engine

3 Experiments

In this section we present the SMT system and the training and test data used in the experiments.

3.1 The SMT System

The SMT system follows the description of the baseline architecture given for the EMNLP 2011 Sixth Workshop on SMT[7]. The system uses Moses[8], an SMT system that allows the user to automatically train translation models for the language pair needed, considering that the user has the necessary parallel aligned corpus. More details about Moses can be found in [8].

While running Moses, we used SRILM - [16] - for building the language model (LM) and GIZA++ - [11] - for obtaining word alignment information. We made two changes to the specifications given at the Workshop on SMT: we left out the tuning step and we changed the order of the language model (LM) from 5 to 3.

Leaving out the tuning step has been motivated by results we obtained in experiments which are not the topic of this paper, when comparing different settings for SMT: not all tests for the system configuration which included tunning showed improvement in the evaluation scores. Changing the LM order has been motivated by results reported in the SMART project, in which it has been concluded that 3-gram configurations provide best results – see [13].

3.2 Training Data

The training data is part of the JRC-Acquis corpus for English-Romanian. JRC-Acquis is a freely available parallel corpus in 22 languages, which consists of European Union (EU) documents of a legal nature. It is based on the Acquis Communautaire (AC), the total body of European Union (EU) law applicable in the EU Member States. This collection of legislative text changes continuously and currently comprises selected texts written between the 1950s and today.

From the two types of sentence alignments available (Vanilla and HunAlign), we used the Vanilla alignments. The same alignments have been also used in [6]. The sentence alignment is done at paragraph level[9], where a paragraph can be a simple or complex sentence, a sub-sentential phrase (e.g. noun phrase) or even more sentences. In order to reduce possible errors, only one-to-one alignments have been considered for the experiments presented in this paper. More details on the JRC-Acquis corpus can be found in [15].

The corpus has not been (manually) corrected. Therefore, translation, alignment or spelling errors can have an influence on the output quality.

For the SMT experiments, from 391324 links in 6557 documents, only 336509 links (the one-to-one alignments) have been considered. Due to the cleaning step of the SMT system[10], the number of one-to-one alignment links used for the LM was reduced to 240219 links for the Translation Model (TM). This represents 61.38 % of the initial corpus. The average sentence length is around 14.5 tokens. (In this paper token means a word, a number or a punctuation sign.)

---

[7] www.statmt.org/wmt11/baseline.html

[8] www.statmt.org/moses/

[9] The tag < p > from the initial HTML files.

[10] In the Moses description, all sentences longer than forty tokens are excluded.

3.3 Test Data

We used test data from three different corpora:

- JRC-Acquis itself (Case A) in-domain data;
- Europarl (Case B) 'similar' data (in-domain, different genre out-of-genre data);
- RoGER (Case C) out-of-domain data.

The first two corpora could be considered in the same domain, as both refer to EU matters, but they are of a different genre: JRC-Acquis contains EU regulations; Europarl is extracted from the literal reports of the debates in the European Parliament. RoGER represents a totally different domain, as it contains text from a manual of an electronic device. The separation of these texts has been done by inspection and intuition.

A: JRC-Acquis The tests were run on parts of the JRC-Acquis, which were not used for training. 897 sentences (three sets of 299 sentences A: Test 1, A: Test 2, A: Test 3) were removed before the training step from the initial corpus, in order to be used as test data. Sentences were removed from different parts of the corpus to ensure a relevant lexical, syntactic and semantic coverage. A: Test 1+2+3 data set contains all the sentences.

The test data has not been cleaned, this means that no length restriction is considered and sentences might be repeated. For example, the paragraph "Article NUMBER" repeats itself 53, 44 and 11 times in A: Test 1, A: Test 2 and A: Test 3, respectively. The data is in-domain data. The average sentence length is around 21 tokens.

B: Europarl The Europarl parallel corpus [7] is extracted from the proceedings of the European Parliament (the literal reports of the debates) dating back to 1996 and contains in its last version twenty-one languages.

We extracted from version 6 of the corpus[11] three different test data sets, each of 299 sentences from the English-Romanian data. As for JRC-Acquis, we extracted the data from different parts of the corpus: from the beginning, middle and the end of the corpus. Small corrections have been done, as sometimes also sentences in other languages have been encountered.

The test data sets from this corpus are: B: Test 1, B: Test 2, B: Test 3 and B: Test 1+2+3. The average sentence length is around 13 tokens. However, for B: Test 1 and B: Test 2 it is around 7.5 and for B: Test 3 it is 24.5. The data is in-domain, but it has a different genre when compared with the training data: the structure and discourse of the text are totally different than the ones of the JRC-Acquis. The text refers to similar matters as the training data: European regulations. We consider these test data sets as 'similar' test data.

C: RoGER In order to analyze the performance of SMT systems to a total different type of text input, we used the RoGER corpus.

RoGER is a parallel corpus, aligned at sentence level. It is domain-restricted, as the texts are from a users' manual of an electronic device. The languages included in

[11] Status: February 2011; http://www.statmt.org/europarl/

the development of this corpus are Romanian, English, German and Russian. The corpus was manually compiled and verified: the translations and the (sentence) alignments were manually corrected. It is not annotated and diacritics are ignored. More about the RoGER corpus can be found in [5].

From the 2333 sentences, we extracted 300 sentences from the middle of the corpus and used them as test data (C: Test). The average sentence length is around 15 tokens. The data is entirely out-of-domain.

4 Evaluation Results

We evaluated our translations using three automatic evaluation metrics: BLEU, NIST and TER. The choice of the metrics is motivated by the (linguistic) resources we had available and the results reported in the literature. Due to lack of data and further translation possibilities, the comparison with only one reference translation is considered in these experiments.

Although criticized, BLEU (bilingual evaluation understudy) is the score mostly used in the last years for MT evaluation. It measures the number of n- grams, of different lengths, of the system output that appear in a set of reference translations. More details about BLEU can be found in [12].

The NIST Score, described in [3], is similar to the BLEU score in that it also uses n-gram co-occurrence precision. If BLEU considers a geometric mean of the n-gram precision, NIST calculates the arithmetic mean. Another difference is that n-gram precisions are weighted by the n-gram frequencies.

TER calculates the minimum number of edits needed to get from obtained translations to the reference translations, normalized by the average length of the references. It considers insertions, deletions, substitutions of single words and an edit-operation which moves sequences of words. More information about TER can be found in [14].

The obtained evaluation results are presented in Tables 2 and 3. The BLEUresults are graphically presented in Fig. 2.

The results for in-domain data are similar to other BLEU scores published in the literature (with the exception of the test data set A: Test 1 for Romanian- English)[12]. The out-of-domain data provides quite low results. The results for ‘similar’ data, somehow surprisingly, are closer to the ones of the out-of-domain data.

A direct comparison with the results in [1] is not possible as there are several important differences, such as the MT approach and the evaluation methodology.

5 Analyzing the Results – Factors of Influence

Several aspects connected with the type of test data can influence the translation results. We will analyze in this paper the number of out-of-vocabulary words

[12] A one-to-one comparison is not possible, as the training and test data are not the same.

Table 2. Evaluation results (Romanian-English)

Test data	BLEU	NIST	TER
A: Test 1	0.2545	3.8325	0.5020
A: Test 2	0.5628	7.6956	0.3756
A: Test 3	0.4271	6.8134	0.4684
A: Test 1+2+3	0.4255	6.9261	0.4457
B: Test 1	0.1372	2.9406	0.9723
B: Test 2	0.1228	3.9758	0.7751
B: Test 3	0.1582	3.6708	0.7562
B: Test 1+2+3	0.1324	4.0559	0.8044
C: Test	0.0621	2.7640	0.7623

Table 3. Evaluation results (English-Romanian)

Test data	BLEU	NIST	TER
A: Test 1	0.3997	6.6279	0.5007
A: Test 2	0.4179	6.8431	0.4898
A: Test 3	0.3797	6.3857	0.5208
A: Test 1+2+3	0.4015	7.4039	0.502
B: Test 1	0.1114	2.7237	0.8315
B: Test 2	0.1057	3.6875	0.7844
B: Test 3	0.1403	3.4697	0.7043
B: Test 1+2+3	0.1128	3.8770	0.7781
C: Test	0.0623	2.7285	0.7340

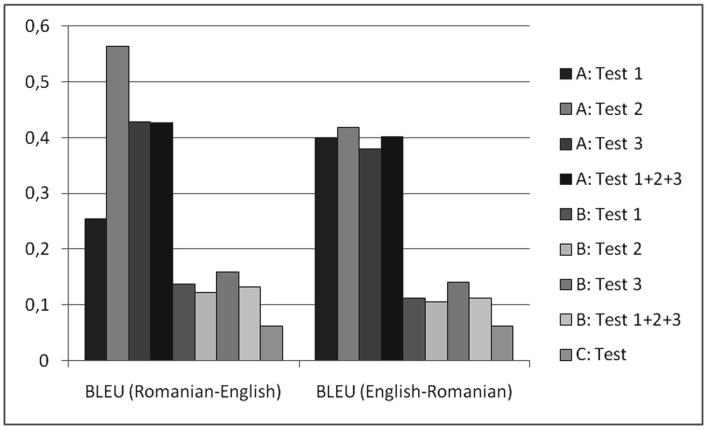

Fig. 2. BLEU results

Table 4. Data description (Romanian-English)

Test data	OOV-words	Sentences in the corpus
A: Test 1	51 (4.10 %)	69 (23.07 %)
A: Test 2	7 (0.76 %)	117 (39.13 %)
A: Test 3	111	81 (27.09 %)
A: Test 1+2+3	169	267 (29.76 %)
B: Test 1	59	0 (0 %)
B: Test 2	697	0 (0 %)
B: Test 3	94	2 (0.66 %)
B: Test 1+2+3	837	2 (0.22 %)
C: Test	330	0 (0 %)

Table 5. Data description (English-Romanian)

Test data	OOV-words	Sentences in the corpus
A: Test 1	33 (3.15 %)	69 (23.07 %)
A: Test 2	2 (0.27 %)	134 (44.81 %)
A: Test 3	96 (8.64 %)	85 (28.42 %)
A: Test 1+2+3	131 (5.59 %)	288 (21.10 %)
B: Test 1	30 (7.5 %)	21 (7.02 %)
B: Test 2	288 (18.68 %)	3 (1 %)
B: Test 3	60 (11.62 %)	22 (7.35 %)
B: Test 1+2+3	366 (17.99 %)	46 (5.12 %)
C: Test	93 (14.65 %)	0 (0 %)

(OOV-words) and test sentences already encountered in the training data. The tests have been run in a realistic scenario, with no human interference (choosing specific sentence average lengths, testing the inclusion in the training data, etc.) on the test data.

The overview of the OOV-words and test sentences already encountered in the training data is presented in Tables 4 and 5. The OOV-words are extracted analyzing only the surface forms. This means that a word can be in the training data as lemma, but a specific word-form[13] might be missing.

Comparing the OOV-words for Test 1+2+3 for Europarl and Test 1+2+3 for JRC-Acquis, we could conclude that these two sets of OOV-words are (al- most) totally different: only three words for English-Romanian and two for Romanian-English are in common in these two sets of OOV-words.

As expected, for the translation direction Romanian-English, the highest number of OOV-words appear in data C (RoGER; out-of-domain data) data (37.67 %). However, for English-Romanian, Test 2 from Europarl (data B; 'similar' data)

[13] Word-form = Declination form, conjugation form, etc.

contains the highest numer of OOV-words: 18.68 %. The out-of-domain data (data C) has only 14.65 % of OOV-words.

A better analysis of the OOV-words in different test data-sets should be made to have a more realistic overview. For example, it could be possible that in data B, due to the text genre, more declination or conjugation forms have been used, when compared with data A. Therefore, the use of a lemmatizer in the translation process could improve the translation results. Concerning the number of test sentences already found in the training data, excluding in-domain data, more sentences have been found for English-Romanian and 'similar' data. For Romanian-English the results for this aspect is similar for both out-of-domain and 'similar' data: under 1 %.

6 Conclusions

In this paper we showed several SMT experiments with different test data (in- domain vs. out-of-domain vs. 'similar' data) using the JRC-Acquis (English- Romanian) corpus for training. The results for in-domain and out-of-domain data are as expected. Somehow surprisingly, the results for 'similar' data are closer to the results for out-of-domain data. The differences in discourse and vocabulary lowered the translation scores for the Europarl tests, although we find ourselves in the same European framework as in the training data. This shows that having only 'similar' data for a specific domain, we cannot always expect good translation results. We can consider the conclusion of this paper limited to the data we used and only as a starting point for further analyses. A manual analysis of the translations should bring a better overview on the automatic scores and the sources of errors. Further experiments with various corpora and language pairs are needed before drawing a final (more general) conclusion

Acknowledgments. Part of the work in this paper was part of the EU-Project ATLAS, supported through the ICT-PSP-Programme of the EU-Commission (Topic "Multilingual Web") and the PhD research conducted by Monica Gavrila at the University of Hamburg (see [4]).

References

1. Calude, A.: Machine translation of various text genres. Presented at 7th Language and Society Conference of the New Zealand Linguistic Society. Hamilton, New Zealand, 12 p., November 2002. (unpublished) (http://www.mt-archive.info/Calude-2003.pdf)
2. Cristea, D.: Romanian language technology and resources go to Europe. Presentation held at the FP7 Language Technology Informative Days, January, 20–11 (2009)
3. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 138–145. Morgan Kaufmann Publishers Inc., San Francisco (2002)
4. Gavrila, M.: improving recombination in a linear EBMT system by use of constraints, Ph.D. thesis, University of Hamburg (2012)

5. Gavrila, M., Elita, N.: Roger - un corpus paralel aliniat. In: Resurse Lingvistice si Instrumente pentru Prelucrarea Limbii Romane Workshop Proceedings, pp. 63–67, Ed. Univ. Alexandru Ioan Cuza, December 2006. Workshop held in November 2006. ISBN: 978-973-703-208-9
6. Ignat, C.: Improving Statistical Alignment and Translation Using Highly Multilin- gual Corpora. Ph.D. thesis, INSA - LGeco- LICIA, Strasbourg, France, 16 June 2009
7. Koehn, P., Europarl: A Parallel Corpus for Statistical Machine Translation, MT Summit (2005)
8. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, pp. 177–180, Prague, Czech Republic, June 2007
9. Koehn, P., Birch, A., Steinberger, R.: 462 Machine Translation Systems forEurope, MT Summit (2009)
10. Niehues, J., Waibel, A.: Domain adaptation in statistical machine translation using factored translation models. In: Proceedings of EAMT, Saint-Raphael (2010)
11. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguistl. **29**(1), 19–51 (2003)
12. Papineni, K., Roukos, S., Ward, T., Zhu, W-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Session: Machine Translation and Evaluation, pp. 311–318. Association for Computational Linguistics Morristown, Philadelphia (2002)
13. Rousu, J., SMART Project: Workpackage 3 advanced language models. Report of the EU project: SMART (2008)
14. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul. J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223–231, August 2006
15. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 2142–2147, May, Genoa, Italy (2006)
16. Stolcke, A.: SRILM - An extensible language modeling toolkit. In: Proceedings of the International Conference on Spoken Language (2002)