Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 541–546
ISBN 978-83-60810-27-9
ISSN 1896-7094

TREF – TRanslation Enhancement Framework for Japanese-English

Bartholomäus Wloka, Werner Winiwarter

***Abstract*—We present a method for improving existing statistical machine translation methods using an knowledge-base compiled from a bilingual corpus as well as sequence alignment and pattern matching techniques from the area of machine learning and bioinformatics. An alignment algorithm identifies similar sentences, which are then used to construct a better word order for the translation. Our preliminary test results indicate a significant improvement of the translation quality.**

***Index Terms*—Machine Translation, Syntactical Analysis, Sequence Alignment.**

I. Introduction

MACHINE translation has been an active research area throughout the last 40 years. During this period, many promising concepts were proposed; however, there is still much room for improvement [1]. Especially when translating languages with radically different surface characteristics, as it is the case for Japanese-English, current machine translation techniques tend to produce unsatisfying results. The problems of automated translation between these languages become readily apparent when looking at current Web-based translations, e.g. from `www.excite.co.jp/world/english`, which is shown in Fig. 1. While the translations of short phrases are of reasonable quality, translation systems struggle with long sentences. This is due to the growing complexity of sentences with increasing length and the vast differences in word and subclause order between these langauges. Additionally, the characteristics of the Japanese language pose a great challenge for translation into other languages in general [2], [3]. Those characteristics are:

- two syllabaries and a system of several thousand kanji, i.e. originally Chinese characters with several pronunciations and readings,
- lack of spaces to delimit word boundaries,
- a very high ambiguity in the grammar, as there exist no articles to indicate gender or definiteness,

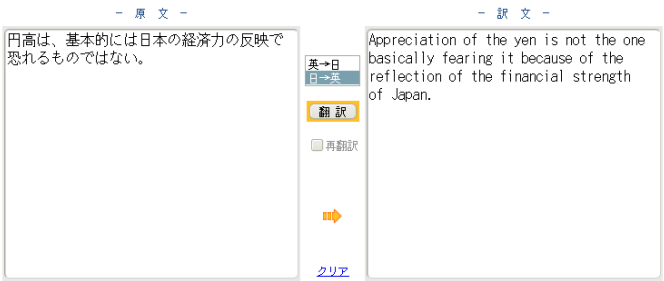

Fig. 1. Example of current Web-based machine translation

- the tendency to omit information which can be inferred implicitly,
- sociolinguistic factors, e.g. avoiding direct and decisive expressions for reasons of politeness,
- an extensive system of formality with several levels of politeness forms, honorific expressions, and humble verb forms depending on the social status, relationship and other factors of the people involved.

To overcome those intricacies, we have directed our attention to a new and interdisciplinary approach. We have designed and implemented a method for finding structurally similar sentences with the help of an algorithm usually employed in the field of bioinformatics [4], [5]. The underlying assumption of our approach is that there is a significant overlap between the **structure** of a sentence and its **meaning**. In this paper, we show that it is possible to enhance statistical machine translation results using this assumption. The *TRanslation Enhancement Framework* (TREF) [6] utilizes aligned and clustered sentence pair data to enhance the output of the statistical machine translation system *Moses* [7].

Though trained for the Japanese-English language pair, the system is modular and flexible. An adjustment or extension to other languages is a matter of changing mere implementation details and adding the language-specific resources, such as lexica, parser, corpora, etc. It is important to mention, however, that our translation framework is specifically designed and well-suited for languages with radically different surface characteristics, e.g. European-Asian language pairs.

The rest of this paper is organized as follows: In Sect. II the research relevant to our work is narrated, before we discuss TREF in Sect. III. Section IV presents our evaluation method and the results, followed by a conclusion and future work in Sect. V.

II. Related Work

The ultimate goal of machine translation, i.e. abolishing language barriers, is presented by [8] in an entertaining narration. This ambitious pursuit of a system which will relieve the lingua franca and enable boundless communication between cultures is not quite yet in the realm of the possible. Nonetheless, research efforts towards this goal have been undertaken. In this section, we outline the research relevant to our work.

A. Corpora

A vital resource for machine translation are bilingual corpora. Unfortunately, these are very rare, especially for the

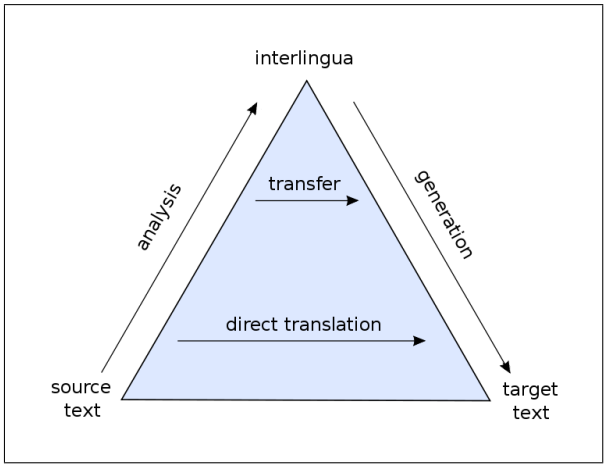

Fig. 2. Translation pyramid

Japanese-English language pair. The currently predominant ones are the *Tanaka corpus* [9], the *Jenaad corpus* [10], and the *Verbmobil treebank* [11]. The Verbmobil treebank contains dialogs from telephone conversations in English, German, Japanese, and other languages, collected during the speech recognition research project of Verbmobil. The Japanese part contains around 160,000 words of text and is written in *Romaji*, i.e. the transcription of Japanese script into Roman literals. The Tanaka corpus consists of roughly 180,000 sentences and has a very broad domain. It has been collected over several years from various sources and compiled by Yasushito Tanaka in 2001. The Jenaad corpus is a collection of close to 150,000 sentence pairs. Extracted from news articles, it offers a certain consistency in terms of sentence types, while still offering a wide range of vocabulary and a variety of grammatical constructs. Because of these qualities, we have chosen the Jenaad corpus for our work. In addition, it is written in Japanese script, thereby avoiding potential ambiguities of the Romaji transcription.

B. Machine Translation

The research in machine translation has ever since included many different approaches. An overview of different techniques can be obtained from [1]. Their visual classification is exemplified by the Vauquois' triangle in Fig. 2 [12]. The historically first method, located at the very top of the triangle, is the *interlingua* approach. It aims towards a language-independent representation, which mediates between two or more languages. In contrast, *statistical machine translation* is at the bottom of the triangle, where no intermediate information is considered in the process, and there is a direct mapping from source to target text, depending on previously trained statistical data. A good overview of this technique can be obtained from [13].

Other approaches, which are also described in more detail in [14], are found somewhere between those two extremes, and the advantage of each depends on the demands of the given language pair. The challenges of translating Japanese to English gave birth to the new idea of *corpus-based* machine translation [15]. Apart from its success in translating between these languages, it further provides the opportunity for enhancing language learning environments by presenting the intermediate steps, i.e. the linguistic analysis of the translation process, to the learner. This was successfully accomplished by [16], [17]. The corpus-based method was quickly adopted by the machine translation community and merged with other techniques, as for example in [18]. Together with the idea of [19], that a mapping of grammatical functions and semantic roles is crucial for the Japanese-English pair, we have decided to mold these ideas into a new approach.

We have chosen a statistical machine translation method for a baseline translation in TREF, since it performs well in terms of translation of individual words and short phrases. It does not adhere to finding transition rules for syntax ordering and therefore leaves a good first candidate for the post-editing done by TREF.

Amongst different tools, we have chosen *Moses*, since it is particularly effective when trained with a sufficiently large bilingual corpus. Moses scores well for structurally similar languages; however, for language pairs like Japanese-English, the word order is disarranged, which significantly lowers the quality of the translation, up to the point where the meaning of the sentence is irrecognizable. Moses does not consider any grammatical rules, so the output is syntactically wrong most of the time. The post-editing and rearranging of the Moses output aims at addressing this problem. Our method finds the correct word order for the translation result and produces a grammatically correct sentence, which conveys the meaning of its English counterpart.

C. Natural Language Processing

To analyze the tokens of our bilingual corpus, we have used the *MontyTagger* from the *MontyLingua* project [20] for English, and *ChaSen* [21] for Japanese. Besides a part-of-speech tagging capability, MontyLingua offers an end-to-end natural language processing toolkit. ChaSen is a high-quality part-of-speech tagger tool for Japanese. Recently, *CaboCha* [22], a Japanese dependency parser, which offers an even wider spectrum of NLP capabilities, has been developed, and we plan to integrate it into TREF in the near future.

D. Sequence Alignment

The Needleman-Wunsch algorithm for computing similarities in protein building blocks, i.e. amino-acid chains, was published in 1970 [23]. Quickly, many derivatives and extensions of this method followed. The basic idea behind this concept was to depict amino-acid chains as strings of alphabetic characters, align them to offer the best match between two strings, and compute a similarity measure [24]. This method was further improved by [25], using a distance measure in conjunction with dynamic programming. Many other research efforts found different distance measures to identify the similarity of sequences. The approach of [26] is generic enough to be extended to the area of machine

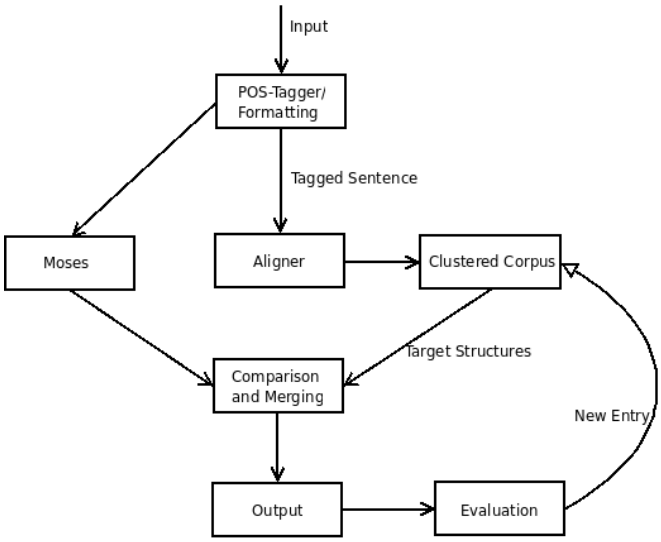

Fig. 3. Overview of dataflow

石炭の利用拡大は大気汚染をさらに悪化させる sekitan no ryou kakudai wa taiki osen wo sarani okka saseru		
石炭/セキタン/2/0/0	の/ノ/71/0/0	利用/リョウ/17/0/0
拡大/カクダイ/17/0/0	は/ハ/65/0/0	大気/タイキ/2/0/0
汚染/オセン/17/0/0	を/ヲ/61/0/0	さらに/サラニ/56/0/0
悪化/アッカ/17/0/0	さ/サ/47/3/5	せる/セル/49/6/1

Fig. 4. Tagged Japanese sentence

expanded use of coal worsens air pollution						
expanded VBN	use NN	of IN	coal NN	worsens VBZ	air NN	pollution NN

Fig. 5. Tagged English sentence

translation, therefore we use it in our research effort by treating sentences from a bilingual corpus analogous to the sequence alignment of amino acid chains.

III. TREF

The overview of the architecture of TREF is shown in Fig. 3. The *PoS-Tagger/Formatting* module tokenizes the input sentence and assigns PoS tags in a format which is described below. The sentences in their tokenized format are then aligned with the clustered corpus to find the target structure, which is sent to the *Comparison and Merging* module. This module takes this input as well as the translation from *Moses* and enhances its translational quality applying a template approach. The resulting translation can then be evaluated and added to the corpus. Each step is described in detail in the following subsections.

A. Part-of-Speech Tagging

The input sentence is sent to either one of the *part-of-speech* (PoS) tagger modules MontyTagger [27] or ChaSen [21]. The result of this process can be seen in Fig. 4 and Fig. 5 for Japanese and English respectively. The Japanese sentence is written in Roman transcription for the reader's convenience. The tags produced by ChaSen consist of a sentence token, its *katakana* representation (one of the Japanese syllabaries, which indicates the pronunciation of a kanji), and a numerical representation of the morphological data. The English tags contain the word itself and the PoS tag as an acronym. After each sentence token is assigned a PoS tag, the sentence and its tags are compared with the sentences already stored in a clustered corpus, which is a customized and enriched version of the *Jenaad* Corpus [10]. We have modified it by removing as much noise as possible, assigned PoS tags to each sentence token, and stored them in an SQL database. We have kept the data with all available PoS tags and additionally created a reduced and optimized tag set, which provides a quick access for efficient processing. Other representations and tag sets can be added easily to satisfy different needs in future work.

B. Aligning and Clustering

In order to identify similar sentences, we have used a slightly modified alignment algorithm from bioinformatics. Instead of aligning protein chains, we align chains of words, i.e. sentences. We have applied relational sequence alignment [4], [5] to obtain clusters of structurally similar sentences. The alignment is done according to the Nienhuys-Cheng distance function.

An example of a distance between the tokens of each sentence is shown in Fig. 6. If the token and its PoS tag differ, the distance is 1. In the case of a structural match, the distance is 0.5, and 0 for a perfect match. The subsequent distance calculation of an entire sentence is depicted in Fig. 7. Gaps, which are identified and symbolized with *(g)* in the example, are assigned variable *gap penalties*. In order to achieve better

d(nn(house),nn(house))	= 0
d(nn(house),nn(office))	= 0.5
d(nn(house),dt(the))	= 1

Fig. 6. Distance calculation example

S1	He	went	to	the	store	to	buy	(g)	milk
T1	PRP	VBD	TO	DT	NN	TO	VB	(g)	NN
S2	She	hurried	to	the	university	to	attend	a	lecture
T2	PRP	VBD	TO	DT	NN	TO	VB	DT	NN
D	0.5	0.5	0	0	0.5	0	0.5	1	0.5

$$SentenceDistance \quad \frac{1}{2 \times 9} \times (0.5+0.5+0+0+0.5+0+0.5+1+0.5) = 0.19444$$

Fig. 7. Sequence alignment distance calculation example

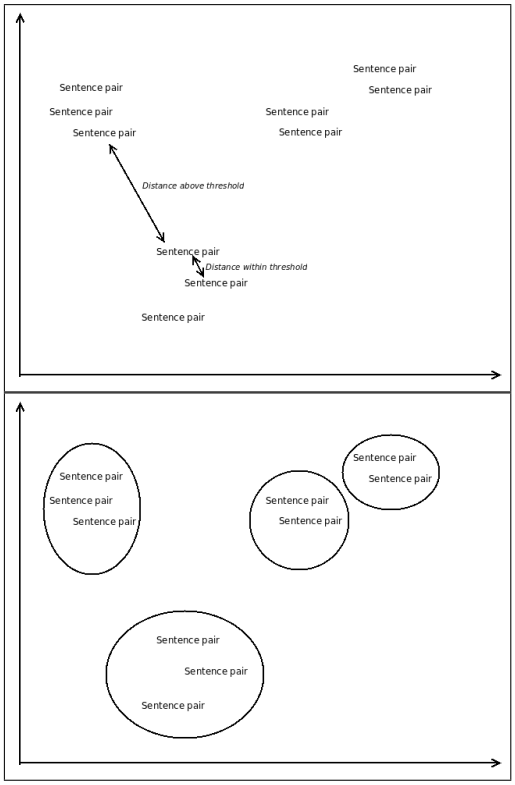

Fig. 8. Clusters in Euclidean space

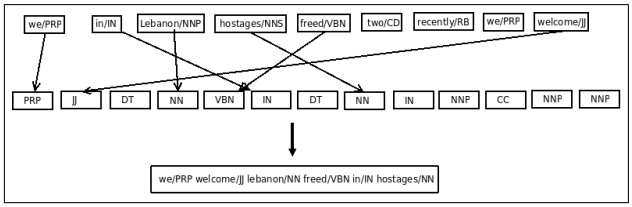

Fig. 9. Matching

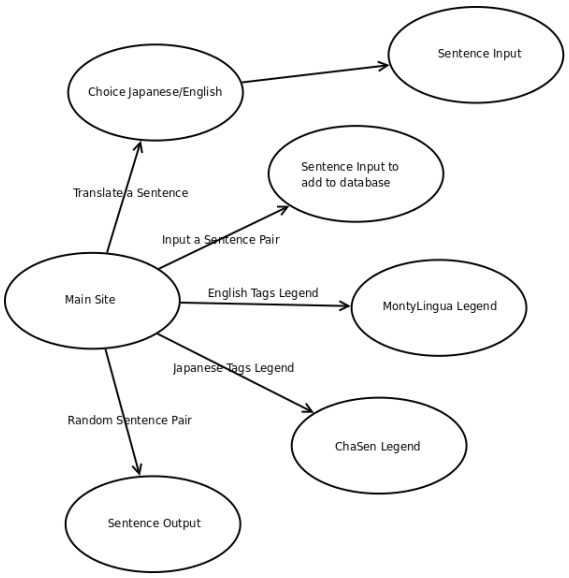

Fig. 10. Structure of the Web framework

matching results, we differentiate between *gap opening* and *gap extension*, which allows us to separate subordinate clauses from otherwise non-matching word sequences.

The similarity measure parameters can be adjusted to fine-tune the result, depending on the text type and text domain. By allowing lower similarity values, a higher number of candidates can be produced, whereas a higher similarity value reduces the number of candidates. This flexibility can be utilized for a language learning application to present an arbitrary amount of similar translations to the student. The output is then evaluated by the user and added to the corpus. Once the distances are computed, clusters can be defined setting a threshold value. This concept is shown in Fig. 8 in a Cartesian coordinate system. Each sentence which has a distance lower than a certain threshold value is assigned to a cluster and is therefore considered *structurally similar* to sentences in this cluster.

C. Comparison and Merging

The comparison of the query sentence with the clusters yields several similar structures. At the same time, the query sentence is processed with Moses to obtain a preliminary translation. This translation is then used to fill the template of the structures which have been found in the previous step. Thereby, a certain number of translation candidates is produced. The filling of the structure templates from the aligning step is shown in Fig. 9. In this example, we use the sentence: “We welcome the progress achieved in the dialog between North and South Korea.” The translation by Moses is: “we in Lebanon hostages freed two recently we welcome”. TREF transforms this by filling the structure template into “we welcome Lebanon freed in hostages”. As can be seen, some tokens are lost in the process of filling the template, which leaves room for future work and potential for further improvement of the translational quality.

D. Web Interface

The clustered corpus of PoS tagged sentence tokens in several representations, as well as morphological information, is stored in a MySQL database and is accessible through a Django Web framework (`http://www.djangoproject.com`). In Django, all interactive content as well as settings, modules, and database setup are written in Python, which made it a good candidate for our system due to its powerful string and text manipulation capabilities. Further, Django provides stable Web development and administrative utilities. In particular, the communication to the database and efficient Web design tools including HTML code inheritance made it an ideal developing environment. The structure of the framework is depicted in Fig. 10. From the main site, the user can navigate to the translation module, the sentence pair input, the random sentence output, as well as legends for the PoS tags for English

and Japanese. The translation module offers an interface, which upon input of a sentence sends it to the server and – after the above described translation process – displays the result. The sentence input module takes a sentence pair input, which is flagged as a new addition and is checked manually before being added to the database. The random sentence output is a first step towards the language learning functionality and outputs a sentence from the database including its translation, its tags, and morphological information. We have created a page for the explanation of PoS tags. The translation of the original Japanese ChaSen tags into English is, to the best of our knowledge, the only English ChaSen PoS-tag legend available.

The framework is available on the Web server maintained by the authors under the URL: `(https://wloka.dac.univie.ac.at/project/)`.

E. Showcase

Figure 11 shows an example of the workflow from the input of a sentence to an output of several translation candidates. The input ”My name is Yamada.” is tagged and compared with the clustered data. The PoS tags for the sentence in this case are: `My/POP` (personal pronoun), `name/NN` (noun), `is/VBZ` (verb), `Yamada/NNP` (proper noun). The alignment detects sentences in the database, which are similar in terms of words and PoS-tags (see Fig. 6). The translations of the identified structures are also checked for similarities within other clusters. This step, which we call *structure-to-meaning-mapping* identifies other structures of potential translation candidates. These structures are sent to the *matching and translation step*, where the structures and the output from Moses are merged to yield the final output, i.e. the translation candidates.

IV. Evaluation

To create a testing scenario, we have extracted 1000 out of the total 150,000 sentences from the Jenaad corpus. The remaining 149,000 sentences were used as training data for Moses and for clustering. Due to the long processing time for each sentence, we have decided to analyze fewer sentences in detail instead of using standard scoring tools, such as [28] or [29], which would be more significant for larger amounts of output. Morover the validity of automated scoring tools of this kind has been criticized by [14], [30]. Hence our evaluation was done by an expert who judged each translation on four categories: word order, word translations, semantics, and fluency. The categories were equally weighted with a top score of 25 each (see Fig. 12). A total of 40 sample sentences were evaluated, and a statistical significance of the result was verified with a Wilcoxon signed-rank test [31]. The result was a better score for the sentences processed with TREF with a score of W=139 over a sample size of N=34 and a P(1-tail) value of 0.119.

V. Conclusion

In this paper, we have described a design for enhancing state-of-the-art machine translation using sequence alignment

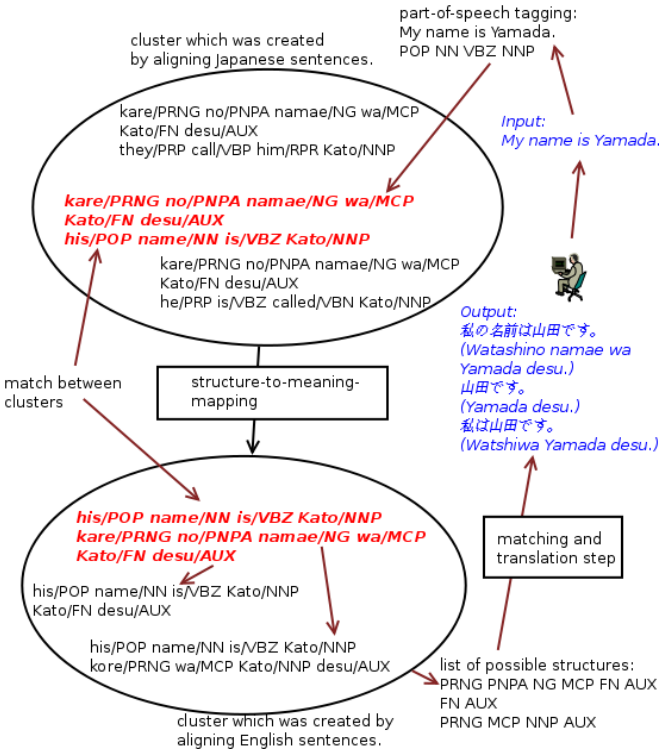

Fig. 11. Translation via Clustering

Input sentence:	我々は、レバノンにおける復興努力を支持する。
Correct Translation:	We support the efforts of reconstruction in Lebanon.
Moses Translation:	we support in lebanon reconstruction efforts .
Enhanced by TREF:	we support lebanon in reconstruction .

	Word Order	Word Translations	Semantics	Fluency
Moses:	5	15	15	5
TREF:	15	15	20	15
Total Score Moses: 40				
Total Score TREF: 65				

Fig. 12. Example evaluation

from the area of bioinformatics, combined with PoS tagging and clustering of a bilingual corpus. Our results have proven that similarities in sentence structure can be used to create templates for translation candidates, in particular for the Japanese-English language pair. We have described our implementation of the system and its Web framework. We have trained the system with the Jenaad Corpus and tested the system for Japanese-English. The evaluation of the system yielded promising results. At the time of writing, TREF is already integrated in another research project focusing on ubiquitous translation and language learning with the help of mobile devices.

For future work, we plan to optimize the parameters in the aligning process to fine-tune the word reordering as well as adding grammatical parsing steps after the template filling to improve the syntactical correctness of the sentence. An

additional dictionary lookup will be integrated to amend word translations, which could not be processed by the statistical translation step.

We want to extend the language learning aspect of the system to offer a Web-based learning platform and improve the efficiency of the entire system with pre-computing and indexing methods. We plan to incorporate a Japanese dependency parser. The currently active research efforts on the Japanese WordNet [32] and CaboCha [22] are promising candidates for an additional extension for TREF as a language learning platform offering extensive semantic and syntactic information as well as visual representations of vocabulary.

REFERENCES

[1] Y. Wilks, *Machine Translation: Its Scope and Limits*. Springer-Verlag, 2008.

[2] Y. McClain, *Handbook of Modern Japanese Grammar*. The Hokuseido Press, 1981.

[3] S. Makino and M. Tsutsui, *A Dictionary of Basic Japanese Grammar*. The Japan Times, 1986.

[4] K. Kersting, L. D. Raedt, B. Gutman, A. Karwath, and N. Landwehr, *Probabilistic Inductive Logic Programming*. Springer Berlin/Heidelberg, 2008, ch. Relational Sequence Learning.

[5] A. Karwath and K. Kersting, "Relational sequence alignments and logos," pp. 290–304, 2007.

[6] B. Wloka, "Enhancing Japanese-English machine translation – a hybrid approach," Master's thesis, University of Freiburg, 2009.

[7] H. Hoang *et al.*, "Moses: Open source toolkit for statistical machine translation," 2007, pp. 177–180.

[8] N. Ostler, Ed., *The Jungle Is Neutral – Newcomer Languages Face New Media*, Foundation for Endangered Languages 172 Bailbrook Lane Bath BA1 7AA England. University Politecnica de Catalunya Barcelona Spain, 2009.

[9] Y. Tanaka, "Compilation of a multilingual parallel corpus," in *Proceedings of the PACLING 2001*, 2001, pp. 265–268.

[10] M. Utiyama and H. Isahara, "Reliable measures for aligning Japanese-English news articles and sentences," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 72–79.

[11] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe-Verbmobil speech recognition engine," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1, p. 83, 1997.

[12] W. J. Hutchins and H. L. Somers, *An Introduction to Machine Translation*. Academic Press, 1992.

[13] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Comput. Linguist.*, vol. 16, no. 2, pp. 79–85, 1990.

[14] C. Boitet, H. Blanchon, M. Seligman, and V. Bellynck, "Evolution of MT with the web," in *Proceedings of the Conference "Machine Translation 25 Years On"*, Cranfield, England, 2009.

[15] M. Nagao, "A framework of a mechanical translation between Japanese and English by analogy principle," in *Proceedings of the international NATO symposium on Artificial and human intelligence*. New York, NY, USA: Elsevier North-Holland, Inc., 1984, pp. 173–180.

[16] W. Winiwarter, "WILLIE – a Web Interface for a Language Learning and Instruction Environment," in *Proceedings of the 6th International Conference on Web-based Learning*. Edinburgh, United Kingdom: Springer-Verlag, 2008.

[17] ——, "WETCAT – Web-Enabled Translation using Corpus-based Acquisition of Transfer rules," in *Proceedings of the Third IEEE International Conference on Innovations in Information Technology*, Dubai, United Arab Emirates, 2006.

[18] M. Carl, A. Way, and W. Daelemans, "Recent advances in example-based machine translation," *Comput. Linguist.*, vol. 30, no. 4, pp. 516–520, 2004.

[19] T. Mitamura and N. Eric, "Hierarchical lexical structure and interpretive mapping in machine translation," in *Proceedings of the 14th Conference on Computational Linguistics*. Morristown NJ USA: Association for Computational Linguistics, 1992, pp. 1254–1258.

[20] H. Liu, "An end-to-end natural language processor with common sense," MIT Media Lab, Tech. Rep., 2004.

[21] Y. Matsumoto, A. Kitauchi, T. Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara, *Japanese Morphological Analysis System ChaSen version 2.2.1*, 2000.

[22] T. Kudo and Y. Matsumoto, "Japanese dependency analysis using cascaded chunking," in *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, 2002, pp. 63–69.

[23] S. Needleman and C. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 2, pp. 443–453, 1970.

[24] T. Smith and M. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981.

[25] A. Karwath and K. Kersting, "Relational sequence alignments," in *Proceedings of the 4th International Workshop on Mining and Learning with Graphs (MLG'06)*, 2006.

[26] A. Karwath, K. Kersting, and N. Landwehr, "Boosting relational sequence alignments," in *Proceedings of the 8th IEEE International Conference on Data Mining*, 2008.

[27] H. Liu and P. Singh, "Conceptnet — a practical commonsense reasoning tool-kit," *BT Technology Journal*, vol. 22, no. 4, pp. 211–226, 2004.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[29] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrency statistics," in *Proceedings of the ARPA Workshop of Human Language Technology*, 2002.

[30] C. Callison-Burch and M. Osborne, "Re-evaluating the role of BLEU in machine translation research," in *Proceedings of the Conference EACL*, 2006, pp. 249–256.

[31] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, December 1945.

[32] F. Bond *et al.*, "Enhancing the Japanese WordNet," in *Proceedings of the 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP*, 2009.

Bartholomaeus Wloka, MSc is a doctoral student at the Department of Scientific Computing, University of Vienna, Austria. He received his BSc degree in 2005 at the University of South Alabama, USA and his MSc degree in 2009 at the University of Freiburg, Germany. His main research interests are human language technology, machine translation and computer-assisted language learning, in particular combined with mobile learning.

Prof. Dr. Werner Winiwarter is the Vice Head of the Department of Scientific Computing, University of Vienna, Austria. He received his MS degree in 1990, his MA degree in 1992, and his PhD degree in 1995, all from the University of Vienna, Austria. The main research interest of Prof. Winiwarter is human language technology, in particular machine translation and computer-assisted language learning. In addition, he also works on data mining and machine learning, Semantic Web, information retrieval, electronic business, and education systems.