WIKIPEDIA

Dictionary-based machine translation

From A to A

Machine translation can use a method based on dictionary entries, which means that the words will be translated as a dictionary does – word by word, usually without much correlation of meaning between them. Dictionary lookups may be done with or without morphological analysis or lemmatisation. While this approach to machine translation is probably the least sophisticated, **dictionary-based machine translation** is ideally suitable for the translation of long lists of phrases on the subsentential (i.e., not a full sentence) level, e.g. inventories or simple catalogs of products and services.[1]

It can also be used to expedite manual translation, if the person carrying it out is fluent in both languages and therefore capable of correcting syntax and grammar.

Contents

LMT

Example-Based & Dictionary-Based Machine Translation

Parallel Text Processing

Lexical Conceptual Structure

"DKvec"

History of Machine Translation

Translingual information retrieval

Machine Translation of Very Close Languages

Multilingual Information Retrieval MLIR

Key words

See also

Bibliography

LMT

LMT[2] is a Prolog-based machine-translation system that works on specially made bilingual dictionaries, such as the Collins English-German (CEG), which have been rewritten in an indexed form which is easily readable by computers. This method uses a structured lexical data base (LDB) in order to correctly identify word categories from the source language, thus constructing a coherent sentence in the target language, based on rudimentary morphological analysis. This system uses "frames"[2] to identify the position a certain word should have, from a syntactical point of view, in a sentence. This "frames"[2] are mapped via language conventions, such as UDICT in the case of English.

In its early (prototype) form LMT[2] uses three lexicons, accessed simultaneously: source, transfer and target, although it is possible to encapsulate this whole information in a single lexicon. The program uses a lexical configuration consisting of two main elements. The first element is a hand-coded lexicon addendum which contains possible incorrect translations.

The second element consist of various bilingual and monolingual dictionaries regarding the two languages which are the source and target languages.

Example-Based & Dictionary-Based Machine Translation

This method of Dictionary-Based Machine translation explores a different paradigm from systems such as LMT. An example-based machine translation system is supplied with only a "sentence-aligned bilingual corpus".[3] Using this data the translating program generates a "word-for-word bilingual dictionary"[3] which is used for further translation.

Whilst this system would generally be regarded as a whole different way of machine translation than Dictionary-Based Machine Translation, it is important to understand the complementing nature of this paradigms. With the combined power inherent in both systems, coupled with the fact that a Dictionary-Based Machine Translation works best with a "word-for-word bilingual dictionary"[3] lists of words it demonstrates the fact that a coupling of this two translation engines would generate a very powerful translation tool that is, besides being semantically accurate, capable of enhancing its own functionalities via perpetual feedback loops.

A system which combines both paradigms in a way similar to what was described in the previous paragraph is the Pangloss Example-Based Machine Translation engine (PanEBMT)[3] machine translation engine. PanEBMT uses a correspondence table between languages to create its corpus. Furthermore, PanEBMT supports multiple incremental operations on its corpus, which facilitates a biased translation used for filtering purposes.

Parallel Text Processing

Douglas Hofstadter through his "Le Ton beau de Marot: In Praise of the Music of Language" proves what a complex task translation is. The author produced and analysed dozens upon dozens of possible translations for an eighteen line French poem, thus revealing complex inner workings of syntax, morphology and meaning.[4] Unlike most translation engines who choose a single translation based on back to back comparison of the texts in both the source and target languages, Douglas Hofstadter’s work prove the inherent level of error which is present in any form of translation, when the meaning of the source text is too detailed or complex. Thus the problem of text alignment and "statistics of language"[4] is brought to attention.

This discrepancies led to Martin Kay’s views on translation and translation engines as a whole. As Kay puts it "More substantial successes in these enterprises will require a sharper image of the world than any that can be made out simply from the statistics of language use" [(page xvii) Parallel Text Processing: Alignment and Use of Translation Corpora].[4] Thus Kay has brought back to light the question of meaning inside language and the distortion of meaning through processes of translation.

Lexical Conceptual Structure

One of the possible uses of Dictionary-Based Machine Translation is facilitating "Foreign Language Tutoring" (FLT). This can be achieved by using Machine-Translation technology as well as linguistics, semantics and morphology to produce "Large-Scale Dictionaries"[5] in virtually any given language. Development in lexical semantics and computational linguistics during the time period between 1990 and 1996 made it possible for "natural language processing" (NLP) to flourish, gaining new capabilities, nevertheless benefiting machine translation in general.[5]

"Lexical Conceptual Structure" (LCS) is a representation that is language independent. It is mostly used in foreign language tutoring, especially in the natural language processing element of FLT. LCS has also proved to be an indispensable tool for machine translation of any kind, such as Dictionary-Based Machine Translation. Overall one of the primary goals of LCS is "to demonstrate that synonymous verb senses share distributional patterns".[5]

"DKvec"

"DKvec is a method for extracting bilingual lexicons, from noisy parallel corpora based on arrival distances of words in noisy parallel corpora". This method has emerged in response to two problems plaguing the statistical extraction of bilingual lexicons: "(1) How can noisy parallel corpora be used? (2) How can non-parallel yet comparable corpora be used?"[6]

The "DKvec" method has proven invaluable for machine translation in general, due to the amazing success it has had in trials conducted on both English – Japanese and English – Chinese noisy parallel corpora. The figures for accuracy "show a 55.35% precision from a small corpus and 89.93% precision from a larger corpus".[6] With such impressive numbers it is safe to assume the immense impact that methods such as "DKvec" has had in the evolution of machine translation in general, especially Dictionary-Based Machine Translation.

Algorithms used for extracting parallel corpora in a bilingual format exploit the following rules in order to achieve a satisfactory accuracy and overall quality:[6]

1. Words have one sense per corpus
2. Words have single translation per corpus
3. No missing translations in the target document
4. Frequencies of bilingual word occurrences are comparable
5. Positions of bilingual word occurrences are comparable

This methods can be used to generate, or to look for, occurrence patterns which in turn are used to produce binary occurrence vectors which are used by the "DKvec" method.

History of Machine Translation

The history of machine translation (MT) starts around the mid 1940s. Machine translations was probably the first time computers were used for non-numerical purposes. Machine translation enjoyed a fierce research interest during the 1950s and 1960s, which was followed by a stagnation until the 1980s.[7] After the 1980s, machine translation became mainstream again, enjoying an even bigger popularity than in the 1950s and 1960s as well as rapid expansion, largely based on the text corpora approach.

The basic concept of machine translation can be traced back to the 17th century in the speculations surrounding "universal languages and mechanical dictionaries".[7] The first true practical machine translation suggestions were made in 1933 by Georges Artsrouni in France and Petr Trojanskij in Russia. Both had patented machines that they believed could be used for translating meaning from a language to another. "In June 1952, the first MT conference was convened at MIT by Yehoshua Bar-Hillel".[7] On 7 January 1954 a Machine Translation convention in New York, sponsored by IBM, served at popularizing the field. The conventions popularity came from the translation of short English sentences into Russian. This engineering feat mesmerised the public and the governments of both the USA and USSR who therefore stimulated large-scale funding in machine translation research.[7] Although the enthusiasm for machine translation was extremely high, technical and knowledge limitations led to disillusions regarding what machine translation was actually capable of doing, at least at that time. Thus machine translation lost in popularity until the 1980s, when advances in linguistics and technology helped revitalise the interest in this field.

Translingual information retrieval

"Translingual information retrieval (TLIR) consists of providing a query in one language and searching document collections in one or more different languages". Most methods of TLIR can be quantified into two categories, namely statistical-IR approaches and query translation. Machine translation based TLIR works in one of two ways. Either the query is translated in the target language, or the original query is used to search while the collection of possible results is translated in the query language and used for cross-reference. Both methods have pros and cons, namely:[8]

- Translation Accuracy – the correctness of any machine translation, is dependent on the size of the translated text, thus short texts or words may suffer from a bigger degree of semantic errors, as well as lexical ambiguities, whereas a larger text may provide context, which helps at disambiguation.
- Retrieval Accuracy – based on the same logic invoked at the previous point, it is preferably to have whole documents translated, rather than queries, because large texts are likely to suffer from less loss of meaning in translation then short queries.
- Practicality – unlike the previous points, translating short queries is the best way to go. This is because it is easy to translate short texts, whilst translating whole libraries is highly resource intensive, plus the volume of such a translating task implies the indexing of the new translated documents

All this points prove the fact that Dictionary-Based machine translation is the most efficient and reliable form of translation when working with TLIR. This is because the process "looks up each query term in a general-purpose bilingual dictionary, and uses all its possible translations."[8]

Machine Translation of Very Close Languages

The examples of RUSLAN, a dictionary-based machine translation system between Czech and Russian and CESILKO, a Czech – Slovak dictionary-based machine translation system, shows that in the case of very close languages simpler translation methods are more efficient, fast and reliable.[9]

The RUSLAN system was made in order to prove the hypotheses that related languages are easier to translate. The system development started in 1985 and was terminated five years later due to lack of further funding. The lessons taught by the RUSLAN experiment are that a transfer-based approach of translation retains its quality regardless of how close the languages are. The main two bottlenecks of "full-fledged transfer-based systems"[9] are complexity and unreliability of syntactic analysis.[10]

Multilingual Information Retrieval MLIR

"Information Retrieval systems rank documents according to statistical similarity measures based on the co-occurrence of terms in queries and documents". The MLIR system was created and optimised in such a way that facilitates dictionary based translation of queries. This is because of the fact that queries tend to be short, a couple of words, which, despite not providing a lot of context it is a more feasible than translating whole documents, due to practical reasons. Despite all this, the MLIR system is highly dependent on a lot of resources such as automated language detection software.[11]

Key words

Linguistics (**lin·guis·tics** | lĭng-gwĭsʹtĭks) = *n. (used with a sing. verb)* The study of the nature, structure, and variation of language, including phonetics, phonology, morphology, syntax, semantics, socio-linguistics, and pragmatics.[12]

computational linguistics = The branch of linguistics in which the techniques of computer science are applied to the analysis and synthesis of language and speech.[13]

Syntax (**sin**-taks) noun = a. the study of the rules for the formation of grammatical sentences in a language; b. the study of the patterns of formation of sentences and phrases from words; c. the rules or patterns so studied; *Computers*. the grammatical rules and structural patterns governing the ordered use of appropriate words and symbols for issuing commands, writing code, etc., in a particular software application or programming language.[14]

See also

- Example-based machine translation
- Rule-based machine translation
- Language industry
- Machine translation
- Statistical machine translation
- Neural machine translation
- Translation

Bibliography

1. Uwe Muegge (2006), "An Excellent Application for Crummy Machine Translation: Automatic Translation of a Large Database", in Elisabeth Gräfe (2006; ed.), *Proceedings of the Annual Conference of the German Society of Technical Communicators*, Stuttgart: tekom, 18-21.
2. Mary S. Neff Michael C. McCord (1990). "ACQUIRING LEXICAL DATA FROM MACHINE-READABLE DICTIONARY RESOURCES FOR MACHINE TRANSLATION". IBM T. J. Watson Research Center, P. O. Box 704, Yorktown Heights, New York 10598: 85–90. CiteSeerX 10.1.1.132.8355 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.132.8355).
3. Ralf D. Brown. "Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation" (http://www.mt-archive.info/TMI-1997-Brown.pdf) (PDF). Language Technologies Institute (Center for Machine Translation) Carnegie Mellon University Pittsburgh, PA 15213-3890 USA. Retrieved 2 November 2015.
4. Jean V´eronis (2001). *Parallel Text Processing: Alignment and Use of Translation Corpora*. *Computational Linguistics*. **27**. Dordrecht: Kluwer Academic Publishers (Text, speech and language technology series, edited by Nancy Ide and Jean V´eronis, volume 13), 2000, xxiii+402 pp; hardbound. pp. 592–595. doi:10.1162/coli.2000.27.4.592 (https://doi.org/10.1162%2Fcoli.2000.27.4.592). ISBN 978-0-7923-6546-4.
5. BONNIE J. DORR (1997). "Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation" (http://download.springer.com/static/pdf/716/art%253A10.1023%252FA%253A1007965530302.pdf?originUrl=http%3A%2F%2Flink.springer.com%2Farticle%2F10.1023%2FA%3A1007965530302&token2=exp=1446652127~acl=%2Fstatic%2Fpdf%2F716%2Fart%25253A10.1023%25252FA%25253A1007965530302.pdf%3ForiginUrl%3Dhttp%253A%252F%252Flink.springer.com%252Farticle%252F10.1023%252FA%253A1007965530302*~hmac=5697c0e4de24aef3be9dcddb7f87900ebac27599d7e34f5d297ed9e55da00dee) (PDF). *Machine Translation*. **12** (4): 271–322. doi:10.1023/A:1007965530302 (https://doi.org/10.1023%2FA%3A1007965530302). Retrieved 2 November 2015.
6. David Farwell Laurie Gerber Eduard Hovy (1998). *Machine Translation and the Information Soup* (http://download.springer.com/static/pdf/42/bok%253A978-3-540-49478-2.pdf?originUrl=http%3A%2F%2Flink.springer.com%2Fbook%2F10.1007%2F3-540-49478-2&token2=exp=1446652324~acl=%2Fstatic%2Fpdf%2F42%2Fbok%25253A978-3-540-49478-2.pdf%3ForiginUrl%3Dhttp%253A%252F%252Flink.springer.com%252Fbook%252F10.1007%252F3-540-49478-2*~hmac=4f84c22e743cb0db2f7e3e3bd46d5136fc975e1a6d780d0cce451a512c1e93d2) (PDF). Lecture Notes in Computer Science. **1529**. CR Subject Classification (1998): I.2.7, H.3, F.4.3, H.5, J.5 Springer-Verlag Berlin Heidelberg New York. doi:10.1007/3-540-49478-2 (https://doi.org/10.1007%2F3-540-49478-2). ISBN 978-3-540-65259-5. Retrieved 2 November 2015.
7. J. Hutchins (January 2006). *Machine Translation: History* (http://www.sciencedirect.com/science/article/pii/B0080448542009378). *Encyclopedia of Language & Linguistics*. pp. 375–383. doi:10.1016/B0-08-044854-2/00937-8 (https://doi.org/10.1016%2FB0-08-044854-2%2F00937-8). ISBN 9780080448541. Retrieved 2 November 2015.

8. Yiming Yang; Jaime G. Carbonell; Ralf D. Brown; Robert E. Frederking (August 1998). "Translingual information retrieval: learning from bilingual corpora" (http://www.sciencedirect.com/science/article/pii/S0004370298000630). *Artificial Intelligence*. Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA. **103** (1–2): 323–345. doi:10.1016/S0004-3702(98)00063-0 (https://doi.org/10.1016%2FS0004-3702%2898%2900063-0). Retrieved 2 November 2015.
9. Jan HAJIC; Jan HRIC; Vladislav KUBON (2000). "Machine translation of very close languages" (http://dl.acm.org/citation.cfm?id=974149). *Proceedings of the sixth conference on Applied natural language processing* -. pp. 7–12. doi:10.3115/974147.974149 (https://doi.org/10.3115%2F974147.974149). Retrieved 2 November 2015.
10. Ari Pirkola (1998). "The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval" (http://dl.acm.org/citation.cfm?id=290957). *The Effects of Query Structure and Dictionary Setups in DictionaryBased Cross-language Information Retrieval*. Department of Information studies University of Tampere. pp. 55–63. CiteSeerX 10.1.1.20.3202 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.3202). doi:10.1145/290941.290957 (https://doi.org/10.1145%2F290941.290957). ISBN 978-1581130157. Retrieved 2 November 2015.
11. David A. Hull; Gregory Grefenstette (1996). "Querying across languages" (http://dl.acm.org/citation.cfm?id=243212). *Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval*. Rank Xerox Research Centre 6 chemin de Maupertuis, 38240 Meylan France. pp. 49–57. doi:10.1145/243199.243212 (https://doi.org/10.1145%2F243199.243212). ISBN 978-0897917926. Retrieved 2 November 2015.
12. "linguistics" (http://www.thefreedictionary.com/linguistics).
13. "computational linguistics - definition of computational linguistics in English from the Oxford dictionary" (http://www.oxforddictionaries.com/definition/english/computational-linguistics). *www.oxforddictionaries.com*. Retrieved 2015-11-04.
14. "The definition of syntax" (http://dictionary.reference.com/browse/syntax). *Dictionary.com*. Retrieved 2015-11-04.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Dictionary-based_machine_translation&oldid=869186104"

This page was last edited on 16 November 2018, at 23:28 (UTC).