Computerlinguistische Analyse mehrsprachiger Fachtexte

Ingeborg Blank
Centrum für Informations- und Sprachverarbeitung
Universität München, Dissertation

Durch die zunehmende internationale Zusammenarbeit und wirtschaftliche Verflechtung ist der Bedarf an Übersetzungen von Fachtexten in den letzten Jahren ständig gestiegen. So ist es verständlich, daß gegenwärtig ein starkes Interesse an effizienten Werkzeugen zur Unterstützung des Übersetzungsprozesses bzw. an Maschineller Übersetzung besteht. Ein multilinguales Textkorpus (eine Sammlung von Texten und deren Übersetzungen in mehrere Sprachen) stellt eine wichtige Informationsquelle für einen Übersetzer, insbesondere bei Terminologiefragen, dar. Ein solches Textkorpus bedarf aber einer entsprechenden Aufbereitung, um seinen vollen Nutzen entfalten zu können. Diese Idee war der Ausgangspunkt für die vorliegende Arbeit. Das Ziel besteht darin, ein Verfahren zu entwickeln, mit dem ein bilinguales Fachtextkorpus maschinell so aufbereitet werden kann, daß die darin enthaltenen lexikalischen bzw. terminologischen Ressourcen explizit verfügbar werden. Die Aufbereitung sollte es ermöglichen,

- Vorkommen möglicher Fachtermini aus den Daten zu extrahieren und
- nach Auffinden eines relevanten Beispiels in einem Quelltext die entsprechende Stelle im Zieltext zu lokalisieren.

Für die Untersuchung standen dreisprachige (dt.-engl.-frz.) Texte aus der Patentdokumentation (insgesamt 12 Mio. Wörter) zur Verfügung. Eine maschinelle Aufbereitung unter terminologischen Gesichtspunkten erfordert

1. eine Parallelisierung des bilingualen Korpus, d. h. eine Alignierung der Texte mit ihren Übersetzungen auf Satzebene,
2. eine Definition von Fachtermini mit formalisierbaren Kriterien, d. h. eine Definition, die in einem automatischen Verfahren umsetzbar ist.

Eine solche Definition liegt beispielsweise dann vor, wenn ein Fachterminus als Nominalphrase mit sprachspezifischen Bildungsmustern definiert wird, z. B. im Deutschen als Nominalkompositum (Beschwerdeverfahren), im Englischen als eine Nomen-Nomen-Verbindung (appeals procedure) und im Französischen als Verbindung aus einem Nomen und einer Präpositionalgruppe (proddure de recours).

3. die Extraktion der Fachtermini, d. h. die automatische Lokalisierung der Fachtermini in den Texten.

Zwei Implementierungen von Satzalignierungsverfahren wurden für alle Sprachpaare des Korpus evaluiert, die höchste Korrektheit lag bei ca. 97 %. Da die Extraktion von Termini eng an die einzelsprachlichen Besonderheiten der Wortbildung gebunden ist, wurde im Rahmen einer Untersuchung, die den Schwerpunkt mehr auf den qualitativen als auf den quantitativen Aspekt legt, die Extraktion von Termini auf das Sprachpaar Deutsch-Französisch begrenzt.

Die Extraktion der Fachtermini erfordert eine Vorverarbeitung der Texte, deren wichtigste Bestandteile die Lemmatisierung (Verfahren, das die Wortformen eines Textes auf eine kanonische Grundform zurückführt, z. B. Verben auf den Infinitiv, Substantive auf die Form des Nominativ Singulars) und das POS-Tagging (Verfahren, das die Wortformen eines Textes mit Wortklassen [engl. *part of speech* bzw. POS] annotiert) sind. Nach der eigentlichen Extraktion muß geprüft werden, ob die für jede Einzelsprache extrahierten Sequenzen linguistisch korrekte Nominalphrasen sind, um in einem zweiten Schritt der Frage nachzugehen, ob es sich inhaltlich und funktional um Fachterminologie handelt. Diese zweite Frage kann letztlich nur von einem Terminologen beantwortet werden. Eine Auswertung der extrahierten Sequenzen nach statistischen Kriterien kann allenfalls Hinweise auf mögliche Fachtermini geben.

Auf der Grundlage von Literatur aus der Terminologielehre, der Übersetzungswissenschaft und der Computerlinguistik wurden potentielle Fachtermini im Deutschen und im Französischen als Nominalphrasen bestimmter Bildungsmuster definiert. Die Vorverarbeitung und die Extraktion wurden für das Französische mit dem INTEX-System (Silberztein 1993) und eigenen Programmen durchgeführt; bei den deutschen Texten erfolgte die Vorverarbeitung mit dem CISLEX-System (Maier-Meyer 1995) und die Extraktion wurde mit eigenen Programmen durchgeführt. Im Vorfeld wurde die Eignung der beiden Systeme für die jeweilige Aufgabe untersucht und, soweit erforderlich, eine Anpassung durchgeführt.

Im Französischen waren ca. 80 % der extrahierten Bildungsmuster linguistisch korrekte Nominalphrasen. Fehler in der Extraktion sind zum größten Teil auf inkorrektes POS-Tagging und auf Ambiguitäten bei der Bestimmung der Bildungsmuster zurückzuführen. Letztere sind nur durch eine intellektuelle Bearbeitung zu beheben. Im Deutschen waren ca. 95 % der extrahierten einfachen Nomina, Nominalkomposita und Adjektiv-Nomen-Verbindungen linguistisch korrekte Nominalphrasen. Weitere Bildungsmuster konnten nur mit einer Korrektheit von 65 % extrahiert werden; dieses Ergebnis kann maschinell nur durch eine zumindest partielle syntaktische Analyse der Texte behoben werden.

Als mögliche Anwendung wird ein Konkordanzprogramm vorgestellt, das drei Aufgaben erfüllt:

- Lokalisierung von Vorkommen einzelner Fachtermini im Quell- und im Zieltext für die Ermittlung von Übersetzungsäquivalenten
- Analyse spezieller Verwendungskontexte einzelner Fachtermini (Präpositionen, Verben, Adjektive usw.)
- Zusammenstellung von Termini mit gemeinsamen Konstituenten in einer netzwerkartigen Struktur.

Die Ergebnisse dieser Arbeit können für die Erstellung mehrsprachiger Glossare und Terminologiedatenbanken sowie für die Entwicklung eines Übersetzungsspeichers (translation memory) genutzt werden. Mehrsprachige Terminologie wird auch bei der Entwicklung und Anpassung von Systemen zur Maschinellen Übersetzung für neue Fachgebiete benötigt.

Auswahlbibliographie

Arntz, Reiner; Pieht, Heribert: Einführung in die übersetzungsbezogene Terminologiearbeit, Hildesheim: Olms, 1982.

Bourigault, Didier: LEXTER, un Logiciel d'EXtraction de TERminologie: Application à l'acquisition des connaissances à partir de textes. Diss. Ecole des Hautes Etudes en Sciences Sociales, Paris, 1994.

Chureh, Kenneth W.; Dagan, Ido: Termight: Identifying and translating Technical Terminology. In: Proc. of the 4th Conference on Applied Natural Language Processing, Stuttgart, 1994, 5. 34–40. Daille, Béatrice: Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres lin-

guistiques. Thèse de doctorat en informatique fondamentale, Université Paris VII, 1994.

Eijk, Pim van der: Automating the acquisition of Bilingual Terminology. In: Proc. of the Meeting of the European Chapter of the Association for Computational Linguistics, Utrecht, 1993, S. 113–119.

Felber, Helmut; Budin, Gerhard: Terminologie in Theorie und Praxis, Tübingen: Narr, 1989.

Fluek, Hans-Rüdiger: Fachsprachen, Tübingen: Francke, 3. Auflage, 1985.

Gale, William A.; Chureh, Kenneth W.: A program for aligning sentences in bilingual corpora. In: Proc. of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, 1991.

Gaussier, Eric: Extraction automatique de lexiques bilingues par des méthodes statistiques. Thèse de doctorat en informatique fondamentale, Université Paris VII, 1995.

Isabelle, Pierre: Bi-textual aids for translators. In: Proc. of the Annual Conference of the UW Center for the New OED and Text Research, 1992.

Kay, Martin; Röscheisen, Martin: Text-Translation Alignment. In: Computational Linguistics, Volume 19, Number 1, 1993, 5. 121–142.

Krause, Jürgen (Hrsg.): Inhaltserschließung von Massentexten, Hildesheim: Olms, 1988.

Maier-Meyer, Petra: Lexikon und automatische Lemmatisierung. Centrum für Informations- und Sprachverarbeitung (CIS-Bericht 95-84), München, 1995.

Reinart, Sylvia: Terminologie und Einzelsprache, Frankfurt: Lang, 1993.

Silberztein, Max: Dictionnaires éléctroniques et reconnaissance lexicale automatique, Paris: Masson, 1993.

Sta, Jean-David: Comportement statistique des termes et acquisition terminologique à partir de corpus. In: Traitement automatique des langues, Vol. 36, 1995, 5. 119–132.

Waliner, Margot: Untersuchung von Text-Alignment im Rahmen computergestützter Textübersetzung auf der Basis des Kay-Röscheisen-Algorithmus. Diplomarbeit Fachhochschule München, Fachbereich Informatik, 1994.