MARSILIUS KOLLEG

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Automatische Übersetzung zwischen Bild und Text: Visuelle und textuelle Phrasen

Stefan Riezler

Auszug aus dem Jahresbericht
„Marsilius-Kolleg 2015/2016“

Stefan Riezler

Automatische Übersetzung zwischen Bild und Text: Visuelle und textuelle Phrasen

Erwin Panofsky, einer der bedeutendsten Kunsthistoriker des 20. Jahrhunderts, verwendet das Beispiel des „Hutziehens“ zur Verdeutlichung verschiedener Bedeutungsebenen in Kunstwerken im Gegensatz zu ihrer Form: Die Konfiguration eines Objekts (Herr) und eines Ereignisses (Hutziehen) konstituiert in der abendländischen Welt den Akt des Grüßens, dessen Bedeutung über die Bedeutung der komponierten Objekte und Ereignisse hinausgeht.[1] Das selbe Prinzip der Komposition neuer Bedeutungseinheiten von „textuellen Phrasen“ aus elementaren Worteinheiten findet sich in der automatischen Sprachübersetzung[2] und auch in der automatischen Bildverarbeitung als „visuelle Phrasen“[3]. In beiden Fällen verändert sich die Bedeutung bzw. das Erscheinungsbild der komponierten Objekte in der Komposition. Beispielsweise kann die Übersetzung des Idioms „kick the bucket“ nicht Wort für Wort geschehen, sondern nur die Komposition betreffen, und entspricht wiederum einer Komposition wie „ins Gras beißen“. In der Bildverarbeitung wurde erkannt, dass die Erkennung eines komponierten Objekts „person riding a horse“ jeweils auf einem bestimmten Erscheinungsbild der komponierten Objekte beruht (hier die Haltung der Person bzw. des Pferdes), sodass eine Erkennung des

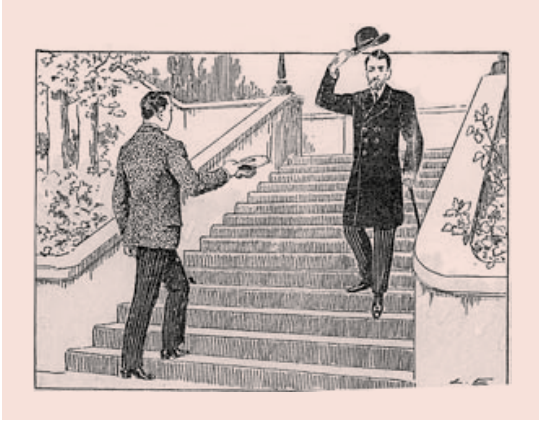

Abb. 1: Visuelle Phrase „Herr zieht Hut“

Kompositums nicht einfach aus der unabhängigen Erkennung der partizipierenden Objekte und der Ermittlung deren Relation erfolgen kann. In diesem Sinne kann auch die Komposition des Objekts (Herr) und des Ereignisses (Hutziehen) als visuelle Phrase angesehen werden, deren kompositionelle Bedeutung für das Grüßen steht (Abb. 1).

Die Forschungsfrage, die von unserer interdisziplinäre Arbeitsgruppe als Marsilius-Kollegiaten untersucht wurde, lautete „Komposition von Bedeutung in Bild und Text" und bestand in der Auslotung der Möglichkeiten einer konkreten Anwendung der Idee der „visuellen Phrasen" mittels Methoden der automatischer Bildverarbeitung (Björn Ommer) und der automatischer Sprachübersetzung (Stefan Riezler) zur Analyse, Verarbeitung und Suche von Bilddaten aus dem Bereich der Kunstgeschichte (Peter Schmidt). Aus der Perspektive der Computerlinguistik, insbesondere der automatischen Sprachübersetzung, wurden von mir und meiner Studentin Mareike Hartmann[4] die folgenden Fragen untersucht:

Können Korrespondenzen zwischen visuellen und textuellen Phrasen aus unbearbeiteten Rohdaten, die in großer Menge verfügbar sind, erlernt werden?

Diese Frage zielt auf mögliche Anwendungen wie die automatische Erkennung visueller Phrasen und Annotierung mit textuellen Phrasen zum Beispiel in digitalisierten Katalogen von Kunstwerke ab. Klarerweise wäre eine manuelle Annotierung von visuellen Phrasen mit textuellen Phrasen eine gute Ausgangsbasis für Methoden maschinellen Lernens, jedoch leider zu aufwendig, um in großem Rahmen angewandt werden zu können. Nehmen wir das Beispiel der visuellen Phrase einer männlichen Gestalt mit einem Messer, die in Kunstwerken bestimmter Epochen mit der textuellen Phrase des Heiligen Georg verbunden ist. Die Rohdaten wären in diesem Fall digitalisierte Kunstwerke mit Abbildungen des Heiligen Georgs und dazugehörigen Bildbeschreibungen, jedoch noch ohne genaue Lokalisierung der visuellen Phrase im Bild, ohne Identifikation der textuellen Phrase in der Bildbeschreibung und ohne Annotation deren Korrespondenz. Der in unserem Projekt verfolgte Ansatz verbindet nun Methoden der automatischen Bildverarbeitung, die es erlauben, visuelle Phrasen automatisch zu erkennen, mit den Mittel der statistischen maschinellen Sprachübersetzung, die verwendet werden, um textuelle Phrasen automatisch zu extrahieren und sie den visuellen Phrasen zuzuordnen.

Können Methoden der statistischen maschinellen Übersetzung auf das Problem der automatischen Übersetzung von textuellen in visuelle Phrasen (und umgekehrt) zugeschnitten werden?

Diese Frage betrifft die konkrete Adaptierung von Techniken der statistischen Sprachübersetzung auf die Sprachen des Bilds und der Bildbeschreibung. In unserem Projekt wurde der Ansatz verfolgt, im Bild visuelle Worte durch Bildverarbeitungsverfahren zu identifizieren und zu repräsentieren, und auf diese Techniken zur Extraktion von bilingualen Phrasen aus dem Bereich der statistischen Sprachübersetzung anzuwenden, mit dem Ergebnis statistischer Korrespondenzen visueller Phrasen und textueller Phrasen. Aufgrund der kleinen Datenmengen von digitalisierten Kunstwerken mit Bildbeschreibungen wurde in unserem Projekt ein Standarddatenset aus der automatischen Bildverarbeitung verwendet, das kurze Bildbeschreibungen (sogenannte „captions") mit einer großen Anzahl von Bildern aus einer begrenzten Menge von Kategorien verbindet. Auf diesen Daten ließen sich suffiziente Statistiken zur Einschätzung der Parameter statistischer Übersetzungsmodelle von visuellen nach textuellen Phrasen erfolgreich erlernen. Somit konnte ein „proof of concept" vorgestellt werden, der bei ausreichender Datenmenge auch auf andere Domänen, insbesondere digitalisierte Kunstkataloge, angewandt werden kann.

Können Paare von visuellen/textuellen Phrasen zu verbesserter automatischer Suche in Bilddatenbanken verwendet werden?

Die letzte Frage unseres Projekts befasste sich mit einer Pilotstudie, die die Ideen der automatischen Übersetzung zwischen visuellen und textuellen Phrasen anhand des konkreten Problems von textbasierter Suche in einer Bilddatenbank evaluieren sollte. Unser experimentelles Design verglich die Zuordnung von Termen der Suchanfrage zu Bildelementen auf der Ebene von Worten (visuelle Worte zu textuellen Worten) mit der Zuordnung von textuellen Phrasen zu visuellen Phrasen. Ein Beispiel sind die Suchanfragen „person carrying surfboard" versus „person on surfboard", wobei nur letztere der visuellen Phrase eines „Surfers" entspricht. In der erfolgen Evaluation war die Genauigkeit in der phrasenbasierten Zuordnung von Bildern von Surfern deutlich höher in der wortbasierten Suchmethode. Zur Illustration sollen unten anstehende Beispiele dienen, die zeigen, dass das besondere Erscheinungsbild von Personen in der Komposition „person on surfboard" die Unterscheidung zur Phrase „person carrying surfboard", aber auch zu Personen, die isoliert erkannt werden, erleichtert.

Abb. 2: Visuelle Phrasen korrespondierend zu textellen Phrasen „person carrying surfboard“ (links) und „person on surfboard“ (rechts)

Fazit

Das Jahr als Marsiliusfellow bot mir die Chance, über den Tellerrand des Computerlinguisten zu schauen und Einblicke in die Arbeitsweise, Methoden und Wissenschaftsphilosophie von Forschungsfeldern wie Psychologie, Psychiatrie, Kunst- oder Wirtschaftswissenschaften zu erlangen. Dabei war für mich einerseits die Tragfähigkeit von Diskussionen aufbauend auf gemeinsamen methodologischen Fundamenten erstaunlich, und anderseits die Fruchtbarkeit des Einnehmens einer anderen Perspektive auf die eigene Arbeit erfreulich. Die wöchentlichen Treffen führten nicht nur zu neuen Einsichten in wissenschaftlicher Perspektive, sondern auch zu professioneller Kollaboration und zu persönlichen Freundschaften. Meine eigene Arbeit wurde nachhaltig vom Thema multimodaler Sprachverarbeitung beeinflusst und resultierte in weiteren Experimenten zur Verbindung von Bild und Text mittels Methoden maschineller Sprachverarbeitung.[5]

1 Vgl. Erwin Panofsky: *Ikonographie und Ikonologie. Eine Einführung in die Kunst der Renaissance*, in: Sinn und Deutung in der bildenden Kunst, S. 36 – 50, Köln: Dumont 1975.

2 Vgl. Philipp Koehn: *Statistical Machine Translation,* Cambridge University Press 2010.

3 Vgl. Mohammad Amin Sadeghi und Ali Farhadi: *Recognition Using Visual Phrases,* in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

4 Mareike Hartmann: *Statistical Machine Translation for Alignments between Images and Captions,* Masterarbeit, Institut für Computerlinguistik, Universität Heidelberg, Heidelberg 2016..

5 Julian Hitschler, Shigehiko Schamoni und Stefan Riezler: *Multimodal Pivots for Image Caption Translation,* in: Proceedings of the 54th Annual Meeting of the Associaton for Computational Linguistics (ACL), Berlin 2016, arXiv:1601.03916.