

Multimodale Versuche der Alignierung historischer Texte

Wagner, Andreas

wagner@rg.mpg.de

Max-Planck-Institut für europäische Rechtsgeschichte,
Deutschland

Bragagnolo, Manuela

bragagnolo@rg.mpg.de

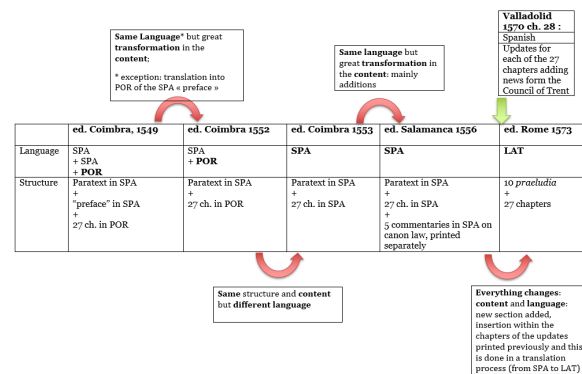
Max-Planck-Institut für europäische Rechtsgeschichte,
Deutschland

Anhand der Aufgabe einer sprachenübergreifenden Kollationierung berichtet dieser Beitrag von "multimodalen" Analysen digitaler Texte: von einer statistischen über lexikalische bis zu wissensmodellierenden Perspektiven auf den Datensatz. Wir greifen auf diese Ansätze zurück, um verschiedene Überarbeitungsstufen und Übersetzungen eines Textes zu alignieren, und wir diskutieren, warum die Aufgabe noch immer keine in der Praxis zufriedenstellende Lösung gefunden hat. So hilft der Beitrag, eine offene Forschungsfrage der Digital Humanities genauer zu bestimmen.

Das Projekt "Das Beichthandbuch des Martín de Azpilcueta und das Phänomen der Epitomierung" untersucht anhand der Entwicklung eines besonderen Textes und seiner Entwicklung den Wandel normativen Wissens in der Rückkopplung mit diversen Praxiszusammenhängen: Der spanische Kirchenrechtler Martín de Azpilcueta (1492-1586) publizierte 1549 sein "Manual de Confesores y Penitentes" mit Regeln für Verfahren und Beurteilung von Beichten. Die ursprüngliche Publikation erschien auf Portugiesisch, allein zu Azpilcuetas Lebzeiten folgten noch über 60 weitere Editionen, in denen der Autor selbst Übersetzungen und Anpassungen vornahm, etwa um auf Beschlüsse des Konzils von Trient einzugehen, oder um das Werk anderen Adressatenkreisen zu erschließen (vgl. Bragagnolo 2018).

Unser Korpus umfasst zunächst 5 zwischen 1549 und 1573 gedruckte Editionen. Zwei auf portugiesisch: (A) Coimbra 1549, 8°, 720 Seiten umfassend, (B) Coimbra 1552, 8°, 1.000 S.; zwei auf spanisch: (C) Coimbra 1553, 4°, 588 S. und (D) Salamanca 1556, 4°, 813 S.; und auf Latein (E) Rom 1573, 4°, 1.136 S. Wir gehen von drei verschiedenen Transformationsmodi aus: Änderungen des Inhalts innerhalb einer Sprache (A # B, C # D); Übersetzung in eine andere Sprache ohne größere Änderungen des Inhalts (B # C); Übersetzung unter gleichzeitiger Änderung des Inhalts (D # E).

Short description of the editions used in the project, highlighting the transformations



Ein erster Beitrag digitaler Methoden zur Analyse dieser Entwicklungen besteht in der systematischen Alignierung von Texten der verschiedenen Versionen über Modifikationen und Übersetzungen hinweg. Wir diskutieren im Folgenden verschiedene Ansätze der automatischen Alignierung von sogenannten Bitexten und wie diese Ansätze sich in der Konfrontation mit den Besonderheiten des Projekts (historisches Vokabular, Orthographie und Grammatik, publizistische oder typographische Eigenheiten in den Texten, inhaltliche Überarbeitungen in den Übersetzungen usw.) bewähren. Ein wichtiger Gesichtspunkt sind dabei immer auch die Art, der Umfang und die Auswirkungen der nötigen manuellen/intellektuellen Vor- und Nachbereitungen.

Für die Evaluation der verschiedenen Ansätze alignieren wir einen Teil der im Projekt als TEI XML transkribierten Texte in der LERA Umgebung¹ manuell. Da die Texte zum Teil umfangreiche Überarbeitungen enthalten, wird zu sehen sein, ob automatische Methoden der Evaluation (wie Papineni et al. 2002 oder Lin/Och 2004) Verwendung finden können, oder ob doch auf eine manuelle Evaluation zurückgegriffen werden muss (ähnlich Darriba Bilbao et al. 2005).

I. Statistische Modi

Im ersten Teil diskutieren wir Algorithmen, die ausblenden, dass es sich bei unseren Daten um symbolische bzw. sprachliche Ausdrücke handelt. Sie werden gleichsam jeweils als "rohe" Datenmengen verstanden, die auf statistische Weisen vermessen werden können und es werden Übereinstimmungen in den Mustern oder in den Intervall-Längen zwischen spezifischen Datenpunkten gesucht.² Obwohl in allen Fällen bestimmte Besonderheiten des historischen Forschungsgegenstands zu Komplikationen führen, ist die Leistungsfähigkeit dieser Ansätze nicht zu unterschätzen. Denn ihre Unzulänglichkeiten sind weitgehend mit jenen besonderen Zusammenhängen historischer Texte verschränkt, in denen ohnehin manuell nach- oder vorgearbeitet werden muss,

und es ist nicht von vornherein auszuschließen, dass es sich lohnen könnte, mit manuellem Aufwand die Texte besser vorzubereiten, um dann mit diesen Ansätzen sehr gute Ergebnisse erzielen zu können.

1. Für die meisten Methoden der computergestützten Übersetzung (*Machine Translation*) stellt der Satz die grundlegende Einheit der Übersetzung dar und es haben sich eine Reihe von Ansätzen etabliert, die zur Erkennung von Satzkorrelationen in Bitexten allein auf die bloße Satzlänge als eines der besten Maße für die Wahrscheinlichkeit abstellen, mit der ein Satz im zu untersuchenden Dokument die Übersetzung eines Referenz-Satzes aus dem Quell-Dokument ist.³ Die Unterschiede in den typischen Satzlängen zwischen zwei Sprachen schlagen sich offenbar in allen Sätzen eines Dokuments in ähnlicher Weise nieder, so dass sich in zwei Dokumenten die Verhältnisse der Satzlängen zueinander stark ähneln. Fälle, in denen allzu kurze Sätze beim Übersetzen verbunden, oder sehr lange Sätze aufgeteilt werden, werden mit geringerer Genauigkeit erkannt; wie häufig dieser Fall aber vorkommt, hängt von den involvierten Sprachen und Übersetzern ab.
2. Ein zweiter Ansatz aus dem *Machine Translation*-Umfeld sind geometrische Ansätze (vgl. Melamed 1999). Sie basieren auf der Annahme, dass es ausreichend sein müsste, sehr grob markierte "Kandidaten" für Satzkorrespondenzen in die richtige Reihenfolge zu bringen. Mit anderen Worten liegt der Fokus nicht auf der eigentlichen Übereinstimmung, sondern auf der Position im Text: Die Ausgangsannahme ist, dass die zu vergleichenden Texte synchron fortschreiten und der erste Satz im einen Text den ersten Satz im anderen übersetzt, der zweite den zweiten usw. Diese Annahme kann in einem durch den Fortschritt in beiden Texten aufgespannten Koordinatensystem als ansteigende Diagonale repräsentiert werden. In einer geometrischen Betrachtung wird dann versucht, durch Umsortierung der vorgefundenen Sätze, die Punkte an diese Diagonale anzunähern.
3. Da unsere Texte in eng verwandten Sprachen vorliegen – Portugiesisch, Spanisch und Latein –, erscheint es lohnenswert, auch mit Ansätzen, die Übereinstimmungen auf der Ebene von Wortstämmen oder -fragmenten untersuchen, einen Versuch zu unternehmen (vgl. Darriba Bilbao et al. 2005). Wir untersuchen also Ähnlichkeiten in den Vektorräumen für die vorkommenden 3- und 4-Gramme.⁴

II. Lexikalische Modi

Eine zweite Menge von Methoden der *Machine Translation* verarbeitet die Daten nur in sprachlogisch aufbereiteter Form, hebt insbesondere auf die übereinstimmende Bedeutung der sprachlichen Ausdrücke

ab und setzt viele "klassische DH"-Ansätze ein (vgl. Ma 2006). Diese Ansätze setzen Arbeitsschritte wie Tokenisierung und Stemmatisierung oder Lemmatisierung voraus und in unseren Experimenten evaluieren wir verschiedene weitere, optionale Schritte, um zunächst zu einer treffenden Charakterisierung *eines* Textes zu gelangen. Dies wird für beide Sprachversionen vorgenommen, bevor dann diese "konzentrierten" oder "gefilterten" Charakterisierungen endlich auf der Basis eines Wörterbuchs *miteinander* verglichen werden.⁵

Die von uns evaluierten optionalen Schritte zur Etablierung einer Charakteristik von Sätzen sind (a) "Filter" wie Stopwörter und TF/IDF-Topwerte und (b) "Booster" wie stärker gewichtete Zahlen, Zahlwörter und Named Entities. Offenkundig hängt allerdings das Ergebnis der Vergleiche in dieser zweiten Perspektive mindestens ebenso sehr von der Qualität der Wörterbücher wie von der Leistung und der Auswahl der vorgeschalteten "Charakterisierungs"-Algorithmen ab. Daher legen wir ein besonderes Augenmerk auf das relative Gewicht der Qualität des Wörterbuchs und ihrer manuellen Verbesserung auf der einen, des Aufwands und Gewinns beim Einsatzes von Filtern und Boostern auf der anderen Seite.

III. Wissensbasierte Modi

Abschließend stellen wir mit der Graphanalyse eine Perspektive vor, die in aktuellen Diskussionen zur sprachübergreifenden Plagiatserkennung diskutiert wird und eine Modellierung des im Text beschriebenen Wissens unternimmt (vgl. Franco-Salvador/Rosso/Montes-y-Gómez 2016). Anstelle eines Wörterbuchs zur Überbrückung des Sprachunterschieds wird hier ein semantisches Netz – in unserem Beispiel BabelNet (Navigli/Ponzetto 2012) – verwendet, um die Wörter der Texte mit "sprachunabhängigen" Konzepten zu verbinden, die untereinander in taxonomischen, synonymen, kontradiktorischen u.a. Beziehungen stehen. Dabei wird durch die Texte jeweils ein Ausschnitt eines umfassenderen Begriffsgraphen instanziiert, um anschließend die resultierenden Teilgraphen miteinander zu vergleichen. Dies erlaubt die Disambiguierung der verwendeten Wörter und eine differenziertere Vergleichsbasis durch die Einbeziehung des semantischen Kontexts der verglichenen Textpassagen. Die Konstruktion und die Vergleiche der zahlreichen Teilgraphen sind offenkundig rechenintensivere Aufgaben, und im Übrigen setzt der Ansatz ebenfalls Schritte wie Tokenisierung und Lemmatisierung voraus, so dass der mögliche Gewinn in der Vergleichsgenauigkeit hier durch einen höheren Aufwand erkauft wird, der zu einem kaum verminderten Aufwand der Textaufbereitung (z.B. der Normalisierung) hinzu kommt.

Diskussion

Wir diskutieren im Rahmen des Beitrags insbesondere, welche Komplikationen sich in der Arbeit in der Folge von Besonderheiten unseres Fragezusammenhangs und Materials gezeigt haben, und wie diese sich auf die unterschiedlichen Ansätze jeweils auswirken. Als wichtige Faktoren konnten wir historische Orthographie und Abkürzungen, uneindeutige und inkonsistente Interpunktion sowie die Kodierung von Layout-Besonderheiten wie Fußnoten identifizieren und so versuchen wir zu bestimmen, welchen Gewinn entsprechende manuelle Vorarbeiten wie Satzsegmentierung, absatzweise Alignierung, Verbesserung des Wörterbuchs, Normalisierung von Schreibungen und Typographie erzielen können.

Fußnoten

1. Paul Molitor, Jörg Ritter et al.: LERA - Locate, Explore, Retrace and Apprehend complex text variants , eine im Rahmen des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projekts SaDA - Semi-automatische Differenzanalyse von komplexen Textvarianten erstellte Arbeitsumgebung.
2. Im linguistischen Kontext entspricht die Mustersuche des ersten Ansatzes etwa einer n-Gramm-Analyse, und wenn die herausgehobenen Datenpunkte des zweiten Ansatzes Repräsentationen von Interpunktionszeichen sind, entspricht dieser einer Analyse der Satzlängen. Obwohl die Auswahl der verwendeten Maße so durchaus durch sprach- und texttheoretische Überlegungen inspiriert und angeleitet ist, sind die Maße selbst von diesen Motiven doch im Grunde vollkommen unabhängig und könnten in gleicher Weise mit ganz anders gearteten Datenreihen angewandt werden. (In Sankoff/Kruskal 1983 etwa werden Anwendungen der Sequenz-Alignierung in ganz anderen Feldern beschrieben.)
3. Die frühesten Versuche in dieser Richtung wurden wohl im IBM Machine Translation Lab unternommen; vgl. Brown et al. 1990. Klassisch wurde der Algorithmus und der Aufsatz von Gale/Church 1993; aktueller, mit weiteren Methoden kombiniert und auf sog. "low resourced languages" zielend etwa bei Varga et al 2005.
4. Als zusätzliche Dimension haben wir dem Vektorraum die Position des jeweiligen Satzes im Text sowie die Behandlung der benachbarten Sätze hinzugefügt, so dass von zwei Satzpaaren mit gleichen n-Gramm-Häufigkeiten dasjenige den Vorzug erhalten kann, dessen Sätze näher beieinander liegen oder das eine mit den benachbarten Sätzen vergleichbare Verschiebung darstellt.
5. Ansätze wie Kay/Röscheisen 1993, Fung/Church 1994 oder auch Varga et al. 2005 können ein solches Wörterbuch auf der Basis allein der vorliegenden Texte erstellen.

Bibliographie

Manuela Bragagnolo: *Les voyages du droit du Portugal à Rome. Le 'Manual de confessores' de Martín de Azpilcueta (1492-1586) et ses traductions*, (The Travels of Law from Portugal to Rome. Martín de Azpilcueta's 'Manual de confessores' (1492-1586) and its Translations), Max Planck Institute for European Legal History Research Paper Series No. 2018-13 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3287684)

Peter F. Brown / John Cocke / Stephen A. Della Pietra / Vincent J. Della Pietra / Fredrick Jelinek / John D. Lafferty / Robert L. Mercer / Paul S. Roossin: *A statistical approach to machine translation*, in: Computational Linguistics 16 (1990): 79-85, <https://dl.acm.org/citation.cfm?id=92860>.

V.M. Darriba Bilbao / J.G. Pereira Lopes / T. Ildefonso: *Measuring the impact of cognates in parallel text alignment*, in: Proceedings of the Portuguese Conference on Artificial Intelligence (2005): 338-343. DOI: 10.1109/EPIA.2005.341306.

Ábel Elekes / Adrian Englhardt / Martin Schäler / Klemens Böhm: *Toward meaningful notions of similarity in NLP embedded models*, in: International Journal on Digital Libraries (2018), DOI: 10.1007/s00799-018-0237-y.

Samuel Fernando / Mark Stevenson: *A semantic similarity approach to paraphrase detection*, in: Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics (2008), 45-52, <https://pdfs.semanticscholar.org/d020/eb83f03a9f9c97e728355c4a9010fa65d8ef.pdf>.

Marc Franco-Salvador / Paolo Rosso / Manuel Montes-y-Gómez: *A systematic study of knowledge graph analysis for cross-language-plagiarism detection*, in: Information Processing and Management 52 (2016), 550-570. DOI: 10.1016/j.ipm.2015.12.004.

Pascale Fung / Kenneth W. Church: *K-vec: A new approach for aligning parallel texts*, in: Proceedings of the 15th Conference on Computational Linguistics, Vol. 2 (1994), 1096-1102, DOI: 10.3115/991250.991328.

William A. Gale / Kenneth W. Church: *A program for aligning sentences in bilingual corpora*, in: Computational linguistics 19/1 (1993): 75-102, <https://dl.acm.org/citation.cfm?id=972455>.

Martin Kay / Martin Röscheisen: *Text-translation Alignment*, in: Computational Linguistics 19/1 (1993), 121-142.

Tom Kenter / Maarten de Rijke: *Short Text Similarity with Word Embeddings*, in: CKIM '15 Proceedings (2015) DOI: 10.1145/2806416.2806475.

Chin-Yew Lin / Franz Josef Och: *Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-bigram Statistics*, in: Proceedings of the 42nd Annual Meeting on

Association for Computational Linguistics (2004). DOI: 10.3115/1218955.1219032.

Xiaoyi Ma: *Champollion: A robust parallel text sentence aligner*, in: 5th International Conference on Language Resources and Evaluation (LREC) 2006, 489-492.

Helena de Medeiros Caseli / Maria das Graças Volpe Nunes: *Evaluation of sentence alignment methods for brazilian portuguese and english parallel texts*, in: Brazilian Symposium on Artificial Intelligence (SBIA) (2004), 184-193, DOI: 10.1007/978-3-540-28645-5_19.

I. Dan Melamed: *A Portable Algorithm for Mapping Bixtext Correspondence*, in: ACL '98 Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (1997), 305-312, DOI: 10.3115/976909.979656.

I. Dan Melamed: *Bitext Maps and Alignment via Pattern Recognition*, in: Computational Linguistics 25/1 (1999), 107-130, <https://dl.acm.org/citation.cfm?id=973218>.

Robert C. Moore: *Fast and accurate sentence alignment of bilingual corpora*, in: S.D. Richardson (ed.): AMTA 2002. Machine Translation: From Research to Real Users, LNCS 2499 (2002), pp. 135-144. DOI: 10.1007/3-540-45820-4_14.

Roberto Navigli / Simone Paolo Ponzetto: *BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*, in: Artificial Intelligence 193 (2012), 217-250, DOI: 10.1016/j.artint.2012.07.001

Kishore Papineni / Salim Roukos / Todd Ward / Wei-Jing Zhu: *BLEU: A Method for Automatic Evaluation of Machine Translation*, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (2002): 311-318. DOI: 10.3115/1073083.1073135.

Christian Paul / Achim Rettinger / Aditya Mogadala / Craig A. Knoblock / Pedro Szekely: *Efficient graph-based document similarity*, in: **H. Sacks et al. (eds.): ESWC '16 European Semantic Web Conference / LNCS 9678 Lecture Notes in Computer Science** (2016), 334-349, DOI: 10.1007/978-3-319-34129-3_21.

Alexandr Rosen: *In search of the best method for sentence alignment in parallel texts*, in: **R. Garab#k (ed.): Computer treatment of Slavic and East European languages. Third international seminar** (2005), 174-185, <http://utkl.ff.cuni.cz/~rosen/public/slovko05.pdf>.

David Sankoff / Joseph Kruskal: *Time Warps, String Edits, and Macromolecules. The Theory and Practice of Sequence Comparison*. Addison-Wesley (1983).

André Santos / José João Almeida / Nuno Carvalho: *Structural Alignment of plain text books*, in: LREC '12 Proceedings of the Eighth International Conference on Language Resources and Evaluation (2012), 2069-2074, http://www.lrec-conf.org/proceedings/lrec2012/pdf/967_Paper.pdf.

Danica Senii#: *Automatic alignment of bilingual sentences. The case of English and Serbian*. M.A. thesis, Louvain, 2016, <https://dial.uclouvain.be/memoire/ucl/en/object/thesis%3A11186>.

Daniel Stein: *Machine translation: Past, present and future*, in: **Georg Rehm / Felix Sasaki / Daniel Stein / Andreas Witt (eds.): Language technologies for a multilingual Europe, TC3 III**. Language Science Press (2018), pp. 5-17. DOI: 10.5281/zenodo.1291924.

Joseph P. Turian / Luke Shen / I. Dan Melamed: *Evaluation of Machine Translation and its Evaluation*, in: Proceedings of Machine Translation Summit Proceedings of Machine Translation Summit IX (2003), 386-393, <https://nlp.cs.nyu.edu/pubs/papers/turian-summit03eval.pdf>.

Dániel Varga / Péter Halácsy / András Kornai / Viktor Nagy / László Németh / Viktor Trón: *Parallel Corpora for medium density languages [Hunalign]*, in: RANLP '05 Proceedings of Recent Advances in Natural Language Processing (2005), 247-258, <http://kornai.com/Papers/ranlp05parallel.pdf>.

Krzysztof Wo#k / Krzysztof Marasek: *A Sentence Meaning Based Alignment Method for Parallel Text Corpora Preparation*, in: Advances in Intelligent Systems and Computing 275 (2014), 107-114, DOI: 10.1007/978-3-319-05951-8_22.