Abschlussbericht: Kodierung internationaler Institutionen – eine Machbarkeitsstudie anhand von ausgewählten Ländern

Patricia Helmich
Fraunhofer ISI
Breslauer Str. 48
76139 Karlsruhe
patricia.helmich@isi.fraunhofer.de
Tel.: 0721 / 6809-339, Fax -176

April 2018

Einleitung. Im Rahmen einer Machbarkeitsstudie wurden anhand dieses Projektes mögliche Verfahren zur weltweiten Institutionskodierung entwickelt und getestet. Dabei sollte das Verfahren soweit möglich automatisiert ablaufen. Es wurde ein Bottom-Up-Verfahren auf Basis der zu kodierenden Daten entwickelt, wobei die bestehenden distinkten Organisationen zunächst wie einzelne Organisationsinstanzen behandelt werden, welche dann weiter zusammen gruppiert werden, um eine Unterscheidung in eindeutige Organisationsobjekte zu erzielen. Dabei wird eine Zuordnung auf Basis von identischen bzw. ähnlichen Namensvarianten der Organisationen getestet. Parallel hat die Universität Bielefeld einen Top-Down-Ansatz zur weltweiten Kodierung der Institutionen basierend auf Wikidata-Einträgen[1] entwickelt, welcher den Institutionen eine eindeutige Wikidata-ID zuweist.

Goldstandard. Zur Evaluierung der Methode wurde zuerst ein Goldstandard-Datensatz entwickelt, auf dessen Basis die Ergebnisse unserer Methode schlussendlich analysiert werden sollen. Dieser beinhaltet Institutionsdaten für die Länder Schweiz, Großbritannien und Südkorea. Als Testdatensatz wurde eine Stichprobe aus Web of Science für jedes der drei Länder gezogen, welche jeweils 2,5% der distinkten Adressen, mindestens aber 500 distinkte Adressen von Organisationen des Landes aus dem Zeitraum 2012-2016 enthalten. Der Zeitraum wurde dabei auf die letzten fünf Jahre beschränkt, um viele alte und bereits geschlossene Institute herauszufiltern. So entstand ein Datensatz von 2462 Institutionen, welchen jeweils eine eindeutige ID zugeordnet wurde. Dabei wurde für die Kodierung soweit möglich auf Wikidata-IDs zurückgegriffen. Die Institutionen, welche nicht in Wikidata enthalten sind, wurden durch ihre Website identifiziert. Zur Vereinfachung der Erstellung des Testdatensatzes wurde von der Universität Bielefeld ein einfaches Verfahren angewendet, welches, soweit möglich, automatisch den Institutionen eine oder mehrere in Frage kommende Wikidata-IDs zuweist. Diese wurden dann manuell überarbeitet bzw. ergänzt. 78% aller Institutionen wurden eine oder mehrere mögliche Wikidata-IDs automatisch zugewiesen. Von allen Institutionen, für welche automatisch eine oder mehrere mögliche Wikidata-IDs gefunden wurden, konnte für 97% dieser Institutionen die korrekte Wikidata-ID unter den zugewiesenen IDs identifiziert werden. 58% aller automatisch gefundenen Wikidata-IDs sind tatsächlich korrekt zugeordnete IDs. Durch die manuelle Überarbeitung und Ergänzung der Wikidata-IDs wurde eine Abdeckung von 97% aller Institutionen durch Wikidata-IDs erzielt, die restlichen Institutionen konnten größtenteils durch ihre Websites identifiziert werden, bei einigen wenigen Institutionen wurde auf den Eintrag in öffentlichen Firmenverzeichnissen oder bei sehr kleinen Institutionen (z.B. eine Person) z.B. auf E-Mail-Adressen zurückgegriffen.

Bei der manuellen Kodierung gab es teilweise Unsicherheiten bzgl. der Zuordnung von IDs, z.B. bei wenig im Internet verfügbaren Informationen über eine Institution; Informationen über eine Institution, welche lediglich z.B. auf Koreanisch vorliegen; oder bei unklaren Informationen über die Zugehörigkeit/Unabhängigkeit einer Institution zu einer anderen.

Evaluationsverfahren. Für die durch Zuordnung entstehenden Cluster bietet sich eine Evaluierung von jeder möglichen Kombination von jeweils zwei Instituten statt der Evaluierung der Kodierung der

1 https://www.wikidata.org

einzelnen Institute an. Im Goldstandard-Datensatz und bei der automatischen Kodierung entstehen unterschiedliche IDs. Dies ist aber nicht weiter schlimm, da lediglich die Abgrenzung der Cluster und nicht deren Label für uns von Bedeutung ist. Um die verschiedenen Kodierungen auch bei überlappenden und verteilten Clustern auswerten zu können, bietet sich die Evaluierung aller möglichen Kombinationen zweier Institute an. Dabei wird jeder möglichen Kombination ein Label „Zusammengehörigkeit" mit den Werten 0 oder 1 zugewiesen. Eine 1 zeigt an, dass beide Institute dieselbe ID besitzen, 0 verweist auf unterschiedliche IDs. Dieses Zusammengehörigkeitslabel wird dann für die automatische Zuordnung auf dem Goldstandard-Datensatz evaluiert, sodass für alle Kombinationen festgestellt wird, ob das automatische Verfahren gemäß dem Goldstandard die Zugehörigkeit beider Institute zum selben Cluster richtig klassifiziert hat.

Tabelle 1: Precision, Recall und F-Score für den Teilstring aus „FULLADDRESS" sowie den gecleanten Teilstring.

	Teilstring „FULLADDRESS"			Teilstring „FULLADDRESS" gecleant		
Land	**Precision**	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**
CHE	100%	75%	86%	98%	76%	86%
GBR	100%	86%	92%	98%	86%	92%
KOR	100%	95%	98%	96%	96%	96%

Automatisches Zuordnungsverfahren über VantagePoint. Das automatisierte Verfahren zur Zuordnung eindeutiger IDs wurde unter Anwendung der Software VantagePoint[2] entwickelt. Die zu identifizierenden Institute werden dabei durch ihren Wert in der Spalte „FULLADDRESS" in der Tabelle „ADDRESSES" in Web of Science repräsentiert. Wir konzentrieren uns zunächst auf den ersten Teil der „FULLADDRESS", d.h. den Teilstring bis zum ersten Komma, welcher im Idealfall den Namen der Institution enthält, während im restlichen String weitere Informationen z.B. über eine spezifische Fakultät und die Adresse des Instituts gegeben werden. Z.B. verwenden wir hier vom String „Univ Bern, Dept Clin Res, Bern, Switzerland" in „FULLADRESS" lediglich „Univ Bern". Es soll nun zunächst geprüft werden, wie gut diese Auswahl bereits funktioniert. Die Zuweisung aller Institutionen zu einem Cluster basierend auf dem beschriebenen Teilstring erzielt für Schweizer Institute eine Precision von 100%, der Recall liegt bei 75%, insgesamt ergibt sich also ein F-Score von 86% (vgl. Tabelle 1). In Großbritannien liegt die Precision bei 100%, der Recall bei 86%, der F-Score bei 92%. Für die koreanischen Institute zeigen sich die besten Ergebnisse mit einer Precision von 100%, einem Recall von 95% und einem F-Score von 98%. Das einfache Abschneiden der weiteren Informationen nach dem ersten Komma des Strings in „FULLADDRESS" erzielt also schon durchaus gute Ergebnisse. Die sehr hohe Precision weist darauf hin, dass es in diesen Ländern keine Probleme mit unterschiedlichen Instituten, welche denselben Namen tragen, gibt. In einem zweiten Testfall wenden wir das Fuzzy Matching-Verfahren in VantagePoint an. Hierbei wird zunächst ein Stemming-Algorithmus auf die Institutsnamen angewandt. Danach gruppiert das Verfahren Institutsnamen

[2] https://www.thevantagepoint.com/

zusammen, welche zu mindestens 68% bzw. 51% (abhängig von der Länge des Namens) übereinstimmen, und vergibt einen neuen, einheitlichen Namen an alle zusammen gruppierten Institute. So werden zum Beispiel die beiden Einträge „Pohang Univ Sci & Technol POSTECH“ und „Pohang Univ Sci & Technol“ der koreanischen Naturwissenschaftlich-Technischen Universität Pohang zusammengeclustert. Die Evaluierung der zusammen gruppierten IDs erhöht den Recall leicht, jedoch in allen drei Ländern lediglich um maximal einen Prozentpunkt, die Precision sinkt in der Schweiz und in Großbritannien um zwei Prozentpunkte und in Korea um vier Prozentpunkte. Der F-Score ist somit für die Schweiz und für Großbritannien identisch, für Korea sinkt er um zwei Prozentpunkte.

Tabelle 2: Fehler, auf deren Basis false negative – Paare bei der Zusammenordnung des ersten Teils der „FULLADDRESS“ entstehen.

Fehler	CHE	GBR	KOR	Gesamtergebnis
Zugehörigkeit nicht erkennbar	37%	33%	33%	35%
Aussagekräftige(s) Wort(folge) gleich	52%	10%	33%	32%
Zu wenig Info im Teilstring	6%	18%	0%	10%
Sprache/Unterschiedliche Bezeichnung in versch. Ländern	0%	23%	0%	9%
Akronym vs. ausgeschriebene Bezeichnung	3%	7%	26%	9%
Unikrankenhaus-Uni	0%	8%	0%	3%
Andere Schreibweise/Abkürzung	2%	0%	9%	2%

Tabelle 3: Fehler, auf deren Basis false negative – Paare bei der Zusammenordnung des gecleanten ersten Teils der „FULLADDRESS“ (Fuzzy Matching) entstehen.

Fehler	CHE	GBR	KOR	Gesamtergebnis
Zugehörigkeit nicht erkennbar	83%	41%	60%	58%
Unikrankenhaus-Uni	0%	24%	0%	12%
Sprache/Unterschiedliche Bezeichnung in versch. Ländern	0%	17%	0%	9%
Aussagekräftige(s) Wort(folge) gleich	17%	3%	0%	7%
Zu wenig Info im Teilstring	0%	14%	0%	7%
Andere Schreibweise/Abkürzung	0%	0%	30%	5%
Akronym vs. ausgeschriebene Bezeichnung	0%	0%	10%	2%

Im Folgenden analysieren wir die Fehler, aufgrund derer die „FULLADDRESS“-Teilstrings fälschlicherweise nicht zusammengeordnet werden (false negatives), was einen Recall unter 100% verursacht (vgl. Tabelle 2). Da die Precision für alle Länder sehr hoch ist, wird auf die false positives an dieser Stelle nicht weiter eingegangen, da die Anzahl sehr gering ist.

Ein Drittel der false negatives basiert darauf, dass die Zusammengehörigkeit der Institute nicht ohne weiteres erkennbar ist. Dabei kann es sich beispielsweise um eine Abteilung eines Krankenhauses handeln, welche einen eigenständigen Namen besitzt, ein Institut mit eigenem Namen, welches einer Universität angehört oder auch zwei Teilinstitute mit unterschiedlichem Namen, welche demselben

Institut angehören. Aufgrund der unterschiedlichen, sich nicht überschneidenden Namen ist es hier schwierig, die Zusammengehörigkeit auf Basis der vorliegenden Daten zu erkennen. Es sind zusätzliche Informationen notwendig, welche den Link zwischen den beiden Instituten aufzeigen. So muss in diesem Falle auch ein menschlicher Annotator beispielweise über zusätzliches Wissen über die Institutionen verfügen bzw. sich dieses über Google/Wikipedia etc. aneignen, um die Institutionen korrekt zusammenzuordnen. Ein weiteres Drittel der false negatives wird dadurch verursacht, dass es zwar Überschneidungen zwischen den beiden Institutsnamen gibt, diese jedoch nur einen kleineren Teil des angegebenen Institutsnamens ausmachen. So überschneiden sich die Institute „Swiss Fed Inst Technol Lausanne EPFL" und "EPFL SB ISIC LCSA" lediglich im Wort „EPFL" und zeigen daher auf den Gesamtstring gesehen wenig Ähnlichkeit. Jedoch hat das Wort „EPFL" eine hohe inhaltliche Aussagekraft, da es auf die „École polytechnique fédérale de Lausanne" hinweist, um die es sich bei beiden Instituten handelt.

Die weiteren false negatives werden verursacht durch zu wenig Information im „FULLADDRESS"-Teilstring (10%), z.B. „Fac Med", die sprachlich unterschiedliche Bezeichnung (9%), z.B. „ETH (Eidgenössische Technische Hochschule)" vs. „Swiss Fed Inst Technol", Akronym vs. ausgeschriebene Bezeichnung (9%), z.B. „UCLH" vs. "Univ Coll London Hosp", Universität vs. Unikrankenhaus (3%) und Abkürzung/alternative Schreibweise vs. vollständiger Name (2%), z.B. "Pohang Univ Sci & Technol" vs. "POSTECH".

Aus den Werten für false negatives, welche durch zu wenig Information im „FULLADDRESS"-Teilstring (im Durchschnitt 10% aller false negatives) verursacht werden, lässt sich folgern, dass der Teilstring eine optimale Ausgangsbasis für das Institutionenmatching ist, da hier eine sehr hohe Precision für alle Länder erreicht wird und die Anzahl der Institute, welche nicht ausreichend Informationen im Teilstring des Namens enthalten, sehr gering ist.

Von allen Fehlern, welche durch die Fuzzy Matching Methode behoben werden können, sind 90% dem Fehler "Aussagekräftige(s) Wort(folge) gleich" zuzuschreiben. Hierbei kann natürlich auch eine aus mehreren Wörtern bestehende aussagekräftige Sequenz gemeint sein. In Fällen, in denen eine hohe Überschneidung der Wörter in den Bezeichnungen der Institute gegeben ist, diese jedoch nicht komplett übereinstimmen, greift die Fuzzy Matching-Methode und ordnet diese zusammen. Durch das Cleaning können sich jedoch auch wieder neue Fehler ergeben bzw. der gleiche Fehler bestehen bleiben. In Tabelle 3 sind die Fehler dargestellt, auf Basis derer sich false negatives der gecleanten Institutionsbezeichnungen ergeben. Hier dominiert klar die nicht erkennbare Zugehörigkeit (knapp 60%). Bei der Mehrheit lag dieser Fehler auch schon vor dem Cleaning vor, sodass aufgrund fehlender korrekter Zuordnungsmöglichkeiten auf Basis von Überschneidungen kein sinnvolles Cleaning möglich war. In einigen Fällen wurden auch aussagekräftige Wörter in den Institutsbezeichnungen fälschlicherweise herausgefiltert, welche vorher für eine Überschneidung mit anderen zugehörigen Institutsnamen verantwortlich waren.

Um den Recall weiter zu erhöhen, wurden mehrere Experimente durchgeführt, in denen der Threshold für die Überschneidungsquote der Institutsbezeichnungen weiter gesenkt wurde. Dadurch steigt der Recall an, die Precision dagegen sinkt. Um einer stark reduzierten Precision entgegenzuwirken, wurde die Stadt der Institutionen bei der Zuordnung berücksichtigt. Zwei Institutionen können somit nur gematcht werden, wenn sie im Feld "CITY" den gleichen Eintrag

besitzen. Ein Problem hierbei ist allerdings, dass für nur etwa 85% aller zusammengehörenden Institutsgruppen allen darin jeweils enthaltenen Institute die gleiche Stadt zugeordnet ist.

Während durch das Absenken des Thresholds für die Überschneidungsquote der zusammengehörenden Institutsnamen der Recall nur langsam ansteigt, fällt die Precision stark ab. So wird z.B. für die Schweiz bei einem Recall von 82% eine Precision von 58% erzielt, sodass sich ein stark gesunkener F-Score von 68% ergibt. Es stellt sich heraus, dass der zunächst angesetzte Threshold von 68% optimal ist für ein ausgeglichenes Verhältnis von Precision und Recall basierend auf der Überschneidungsrate.

Ein weiterer Anwendungsfall des Fuzzy Matching-Verfahrens ist das Herausfiltern von Rechtschreib- und Tippfehlern sowie leicht unterschiedlichen Schreibvarianten (Bindestrich etc.). Der Nutzen dieses Features wird in einem weiteren Experiment untersucht. Während in den vorherigen Tests die einzelnen Wörter komplett identisch sein mussten, um zum notwendigen Gesamtmatch der Institutsnamen gezählt zu werden, wird nun der Threshold für den Match der einzelnen Wörter heruntergesetzt. Auch hierbei wird wieder der Abgleich der Städte angewandt, um das Absinken der Precision zu verhindern. Für die Schweiz kann somit eine Verbesserung von je einem Prozentpunkt für Precision und Recall erreicht werden im Vergleich zur vorherigen Cleaningmethode ohne Fuzzy Wortmatching; der F-Score steigt dadurch jedoch nur leicht und misst weiterhin etwa 86%. Für Großbritannien ändern sich die Werte nur minimal durch die Anwendung des Fuzzy Wortmatchings, während für Korea bei gleichbleibendem Recall (96%) die Precision auf 94% abfällt und somit ein niedriger F-Score (95%) im Vergleich zu vorherigen Cleaningmethode ohne Fuzzy Wortmatching erzielt wird. Daraus wird deutlich, dass die meisten der einzelnen Wörter in den Institutsnamen keine Schreibvariationen oder Fehler enthalten, sondern größtenteils standardisiert sind.

Zusammenfassend lässt sich also sagen, dass VantagePoint bei Schreibvarianten einzelner Wörter und unterschiedlichen Institutsnamen mit hoher Überschneidung der enthaltenen Wörter gute Ergebnisse erzielen kann. Wir stellen aber fest, dass unterschiedliche Schreibvarianten einzelner Wörter selten vorkommen und sich Institutsnamen häufig nur in einzelnen oder wenigen Wörtern überschneiden. Somit erzielt die automatische Zuordnung über VantagePoint keine signifikanten Verbesserungen der über den Teilstring der „FULLADDRESS“ erzielten Ergebnisse.

Probleme des automatischen Zuordnungsverfahrens und Lösungsansätze. Wenn man das Beispiel „Univ Liverpool“ vs. „Univ Oxford“ betrachtet, fällt auf, dass nicht allein die Anzahl der sich überschneidenden Wörter, sondern auch die Semantik bzw. die Aussagekraft der Wörter bei der Ähnlichkeitsbestimmung eine Rolle spielen. Im ersten Beispiel überschneidet sich eins von jeweils zwei Wörtern der beiden Institute, jedoch ist Univ für sich allein nicht besonders aussagekräftig, denn es bezeichnet lediglich die Art der Institution. Erst durch die Zugabe der Ortsinformation (Liverpool bzw. Oxford) wird ein konkretes Institut benannt und der Unterschied zwischen beiden Instituten wird durch das jeweils zweite Wort markiert. D.h. also, in diesem Beispiel geben die jeweils zwei unterschiedlichen Wörter jeweils einen unterschiedlichen Informationsgehalt an, welcher bei der Ähnlichkeitsbestimmung berücksichtigt werden müsste. In VantagePoint gibt es zwei Möglichkeiten Wörtern verschiedene Gewichtungen zuzuordnen. Einerseits können Wörter, welche nicht zur Ähnlichkeitsbestimmung beitragen, von der Berechnung ausgeschlossen werden. Wenn man z.B. Wörter wie „Univ“ ignoriert, dann können im Beispiel die Institute „~~Univ~~ Liverpool“ und

„~~Univ~~ Oxford“ leicht als unterschiedlich markiert werden. In dem Beispiel „Kantonsspital St Gallen“ vs. „Univ St Gallen“ würde diese Technik jedoch fälschlicherweise zu einer großen Übereinstimmung führen. Eine zweite Methodik in VantagePoint ermöglicht, den einzelnen Wörtern Gewichte zuzuordnen. Die Gewichte bestimmen sich allerdings aus der Position der Wörter im Namen, so kann man z.B. festlegen, dass die ersten Wörter im Namen höher gewichtet werden als die weiter hinten stehenden Wörter. Dieses Verfahren eignet sich wiederum besser für das zweite Beispiel, während es im ersten Beispiel „Univ“ eine hohe Bedeutung bei der Ähnlichkeitsberechnung zuordnet. Noch deutlicher wird es im Beispiel „Swiss Fed Inst Technol Lausanne EPFL” vs. „EPFL SB ISIC LCSA”. Das Wort „EFPL“ bezeichnet eine bestimmte Schweizer Universität, womit es sehr aussagekräftig ist. Da sich beide Institute in genau diesem Wort überschneiden, ist es sehr wahrscheinlich, dass sie auf das gleiche Institut verweisen. Jedoch ist „EPFL“ nur eins von vier bzw. sechs unterschiedlichen Wörtern, wobei es ansonsten keine Überschneidungen zwischen den Institutsnamen gibt, zudem steht „EPFL“ in beiden Namen an völlig entgegengesetzten Positionen. Die Gewichtung, welche in VantagePoint verfügbar ist, ist also in unserem Anwendungsfall nicht effizient anwendbar.

Ein Lösungsansatz für eine korrekte Gewichtung der Wörter für die Bestimmung der Ähnlichkeiten zweier Institutsnamen wäre es, die Gewichtung direkt aus den Daten zu lernen. „Univ“ und „Oxford“ bzw. „Liverpool“ haben einzeln gesehen wenig Aussagekraft. Dementsprechend werden diese Wörter nicht als vollständiger Institutsname vorkommen, d.h. ein Institutsname besteht nicht lediglich aus „Univ“ oder „Oxford“. „EPFL“ dagegen bezeichnet ein spezifisches Institut und wird somit auch als kompletter Institutsname ohne weitere Wörter im Namen zu finden sein. Wenn also ein Wort (z.B. „EPFL“) oder eine Wortfolge (z.B. „Univ Oxford“) häufig als kompletter Institutsname auftreten, so kann dies als ein Indikator für eine hohe Aussagekraft des Wortes bzw. der Wortfolge über die Institutszugehörigkeit eines Instituts gesehen werden, welches das Wort bzw. die Wortfolge als Untermenge seines aus mehreren Wörtern bestehenden Namens (z.B. „*EPFL* SB ISIC LCSA” bzw. „*Univ Oxford* Jesus Coll“) enthält. Diesen Wörtern (bzw. Wortfolgen), welche häufig alleine stehen, wird somit ein höheres Gewicht bei der Ähnlichkeitsberechnung zugewiesen als den restlichen Wörtern, welche immer nur als Untermenge von aus mehreren Wörtern bestehenden Namen auftreten (z.B. „Univ“, „Inst“, „Fac“, “Med“, „Coll“).

Möglicherweise ist diese Gewichtungsmethode nicht nur ein Lösungsansatz für das Problem „Aussagekräftige(s) Wort(folge) gleich“, sondern auch für das Problem der nicht erkennbaren Zugehörigkeit. Es ist z.B. denkbar, dass zwei Wortfolgen „A“ und „B“ als aussagekräftige Institutsnamen identifiziert werden, diese aber nicht zusammengefügt werden, da es keinerlei Überschneidungen der beiden Namen gibt, obgleich diese tatsächlich korrekterweise zusammengehören. Wenn es nun ein Institut gibt, in dessen Namen die beiden aussagekräftigen Institutsbezeichnungen als jeweilige Untermenge identifiziert werden können („A B“), so ist dies ein möglicher Indikator dafür, dass die Institute „A“ und „B“ auf dasselbe Institut verweisen.

Eine weitere in VantagePoint nicht vorhandene, aber zu prüfende Option wäre der Abgleich von Akronym und ausgeschriebenem Namen. Jedoch betrifft dieser eine geringere Anzahl von vorkommenden false negatives.

Manuelle Zuordnung. Um ein optimales Zuordnungsergebnis der Institute zu erhalten, kann in einem zweiten Schritt auch eine manuelle Bearbeitung der Gruppierungen in VantagePoint durchgeführt

werden. Hierbei können Institute bzw. Institutsgruppen weiter zusammen gruppiert werden, sodass der Recall erhöht wird. Ebenfalls können von VantagePoint automatisch erstellte Gruppen wieder getrennt werden, um die Precision zu erhöhen. Dieses kann in VantagePoint einfach durch Drag-and-Drop erfolgen. In unserem Testfall haben wir insgesamt 2462 Institute über alle drei Länder vorliegen. Diese stellen nach dem Goldstandard 570 verschiedene Institute dar, welche dementsprechend zusammengruppiert werden müssen. Durch Nutzung des ersten Teilstrings der „FULLADDRESS“ entstehen 698 unterschiedlich Institutsgruppen, es müssten in diesem Fall also noch einmal knapp 130 manuelle Zuordnungsaktionen durchgeführt werden, um einen maximalen Recall zu erreichen. Außerdem müssen vier Splitaktionen erfolgen, d.h. eine vom VantagePoint erstellte Institutsgruppe enthält zwei oder mehr unterschiedliche Institute. Hierbei ist zu bedenken, dass Splitaktionen einen größeren manuellen Aufwand bedeuten, da hierbei für jedes einzelne enthaltene Institut der Gruppe entschieden werden muss, welcher Institutsgruppe es zugeordnet wird. Die gecleanten Institute beispielsweise ergeben nur 645 Institutsgruppen, somit müssen auch nur 75 manuelle Zuordnungen vorgenommen werden. Jedoch ist in diesem Fall die Precision niedriger und es müssen 26 Splitaktionen erfolgen. Wenn man bedenkt, dass einerseits Splitaktionen mehr Aufwand bedeuten als das Verbinden von Gruppen, und andrerseits bei einer geringeren Precision einer Methode alle einzelnen Institute aller Gruppen betrachtet werden müssen, um herauszufinden, ob eine Gruppe aufgesplittet werden sollte, so ist ein Verfahren mit einer hohen Precision als Basis für die weitere manuelle Verarbeitung sinnvoll.

Der Vorteil der manuellen Weiterverarbeitung der Kodierung in VantagePoint liegen zum einen in der einfachen Handhabung der Erstellung der Gruppierungen sowie des darauf basierenden Thesaurus und zum anderen in einem vergleichsweise korrekten Ergebnis. Ein Nachteil ist der höhere zeitliche Aufwand, wobei dieser im größeren Ausmaß lediglich einmal pro Land durchgeführt werden müsste, da auf Basis dieser manuellen Arbeiten ein Thesaurus für ein Land erstellt werden kann, welcher jedes Jahr für die Neukodierung der Institutsdaten genutzt werden kann und lediglich für neu hinzukommende Institutsnamen ergänzt werden muss.

Zusammenfassung und Ausblick. Die im ersten Teilstring der „FULLADDRESS“ enthaltenen Informationen sind weitgehend ausreichend für die eindeutige Identifikation der Institute und stellen somit eine gute Basis für die Zuordnung der Institutsnamen dar. Eine Zuordnung auf Basis dieser Teilstrings erzielt für alle Länder eine optimale Precision und einen Recall zwischen 75% und 95% für die einzelnen Länder. Die Software VantagePoint eignet sich gut dafür, unterschiedliche Varianten von Namen zu gruppieren, welche sich in einigen Wörtern unterscheiden, aber dennoch eine hohe Überschneidung zeigen. Auch Tipp- und Rechtschreibfehler können herausgefiltert werden. Die untersuchten Institutsnamen enthalten jedoch größtenteils standardisierte bzw. korrekte Wörter. Bei unterschiedlichen Namen für gleiche Institute zeigen sich häufig nur kleine Überschneidungen oder auch komplett unterschiedliche Namen. Somit erzielt die automatische Zuordnung von VantagePoint nur eine sehr geringe Verbesserung der Ergebnisse im Gegensatz zur Zuordnung auf Basis der ungecleanten Teilstrings. Ein möglicher Lösungsansatz wäre die stärkere Gewichtung von aussagekräftigen Wörtern bzw. Wortfolgen, welche Institute eindeutig bezeichnen und häufig in den längeren Namen ebenfalls enthalten sind, um somit den Recall bei der Ähnlichkeitsberechnung weiter zu verbessern. Diese Gewichte könnten auf Basis der Daten gelernt werden.

Außerdem bietet VantagePoint die Möglichkeit, die Zuordnung manuell weiter zu optimieren und darauf basierend einen Thesaurus zu erstellen, welcher jedes Jahr wieder benutzt werden kann für die Kodierung der Institute und lediglich für neu hinzukommende Institutsnamen angepasst werden muss. Durch die Zuordnung der Namen auf Basis der Teilstrings lässt sich bereits ein Großteil der Institute mit einer hohen Precision zuordnen, sodass im Folgenden nur noch ein Teil der Institute bzw. Institutsgruppen weiter gruppiert werden müssen, um den Recall weiter zu erhöhen.

Ein Lösungsansatz, den Recall bei der automatischen Zuordnung weiter zu verbessern, ist es, die Ähnlichkeitsberechnung durch die Gewichtung von Worten und Wortfolgen anzureichern, welche auf Basis der Daten erlernt werden. Alternativ kann dieses möglicherweise auch durch zusätzliche externe Informationen zu den Daten erreicht werden, wie es im Ansatz der Universität Bielefeld untersucht wurde. Dazu sollen die Ergebnisse beider Ansätze noch verglichen werden. Wenn durch den Wikidata-Ansatz verbesserte Ergebnisse erzielt werden können, so könnten diese auch als Basis für die manuelle Überarbeitung in VantagePoint genutzt werden, um einen optimalen Thesaurus bei minimalem manuellen Aufwand zu erzielen.