Leibniz-Gemeinschaft

Abschließender Sachbericht

Titel des Vorhabens:

Overcoming language barriers – Cross-lingual search of bibliographic metadata

Leibniz-Einrichtung: Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)

Aktenzeichen: SAW-2016-ZPID-2

Projektlaufzeit: 01.07.2016 – 30.09.2019

Ansprechpartner: PD Dr. Erich Weichselgartner (wga@leibniz-psychology.org)

Inhaltsverzeichnis

Zusammenfassung

Ziel des Forschungsvorhabens war die empirische Evaluation von vier unterschiedlichen Ansätzen zur Verbesserung des cross-lingualen Information Retrievals (CLIR) anhand der Suchmaschine *PubPsych*. PubPsych wurde 2013 vom ZPID und weiteren Partnern als vertikale Suchmaschine für psychologische Literatur, Tests, Behandlungsprogramme und Forschungsdaten entwickelt. Es enthält keine Volltexte, sondern die bibliographischen Metadaten zu einzelnen Dokumenten, die von den beteiligten Datenpartnern bereitgestellt werden. Jeder Datensatz ist dabei in mindestens einer der vier Sprachen Deutsch, Englisch, Französisch oder Spanisch verfügbar.

Im Rahmen des geförderten Vorhabens wurden vier Ansätze untersucht, mit deren Hilfe eine cross-linguale Suche in PubPsych umgesetzt werden kann: (1) Übersetzung der Suchanfrage; (2) Übersetzung der Metadaten; (3) Englisch als Brückensprache; (4) Wissensbasierter Ansatz. Ziel war es, alle Ansätze im Bezug auf ihre Verbesserung der Suchergebnisse aus Nutzersicht zu evaluieren und den erfolgversprechendsten Ansatz anschließend in PubPsych zu implementieren.

Zunächst wurde der PubPsych-Datenbestand vorbereitet und analysiert, um die geeignetsten Vorgehensweisen bei der Entwicklung der vier Systeme zu definieren und im Vorfeld nötige Anpassungen im Suchmaschinenindex vorzunehmen. Es wurde außerdem ein viersprachiges Lexikon generiert, das Begriffe aus kontrollierten, domänenspezifischen Thesauri (MeSH, APA) und allgemeinen Ressourcen (Wikipedia, Apertium) mitsamt ihren Übersetzungen beinhaltete. Dieses Lexikon wurde zur Übersetzung von kontrolliertem Vokabular in den Metadaten und von Suchanfragen verwendet (Ansätze 1 und 4).

Zur Übersetzung von Titeln und Abstracts der Dokumente (Ansatz 2) wurden Modelle und Verfahren der neuronalen maschinellen Übersetzung (NMT) entwickelt und verwendet. Grundlage für die Optimierung des Modells bildeten neben den bereits in den bibliographischen Records von PubPsych existierenden parallelen Daten weitere speziell für das Projekt zusammengestellte Korpora, die nicht ausschließlich Texte aus dem Bereich der Psychologie beinhalteten. Die Evaluation der NMT-Verfahren geschah mithilfe eines im Rahmen des Projektes erstellten Korpus von 800 psychologischen Abstracts, die auf Satzebene aligniert wurden und in allen vier Projektsprachen vorliegen.

Zum Schluss wurden die vier Ansätze in PubPsych-Testsystemen implementiert. An jedes System wurden dieselben 50 Suchanfragen in allen vier Projektsprachen gestellt. Die von den Systemen als relevant angesehenen Dokumente wurden exportiert und manuell auf ihre Relevanz überprüft. Hierdurch ließ sich derjenige Übersetzungsansatz ermitteln, der die relevantesten Resultate lieferte.

Es zeigte sich, dass Ansatz 2 (Übersetzung der Metadaten) sowohl in der system-orientierten als auch in der nutzer-orientierten Evaluation die besten Werte erzielen konnte. Dieses System erhöht die Menge der gefundenen Dokumente deutlich, da – verglichen mit dem bisherigen System – mehr Dokumente bei einer Suche gefunden werden, deren Metadaten in einer anderen Sprache als der Suchsprache vorliegen. Dies wird durch die Übersetzung von Teilen des Suchmaschinenindex erreicht. Dieser Ansatz wird somit in das Suchportal PubPsych integriert.

Ausgangsfragen und Zielsetzung des Vorhabens

Jedes Jahr werden mehrere Millionen wissenschaftliche Artikel und Ergebnisse in unzähligen Sprachen veröffentlicht. Bedingt durch diese Menge an Informationen und Vielfalt an Sprachen ist es in der wissenschaftlichen Praxis schwer bis unmöglich, sämtliche für den jeweiligen Bereich relevante Literatur zu sichten. In Fachgebieten wie der Psychologie führt dies nicht nur zu potenziell doppelter Forschung, sondern verhindert z.B. auch die Verfügbarkeit von neuem Wissen und somit die Anwendung von neuesten Erkenntnissen in der psychologischen Praxis, wie der Psychotherapie.

Weiterhin beinhaltet Forschung, die nur auf den Erkenntnissen beruht, die in bestimmten Sprachen, wie z.B. Englisch veröffentlicht wurden, überdies das Risiko, dass Schlussfolgerungen nur auf Basis bestimmter Subpopulationen gezogen werden und kulturelle Unterschiede und Vielfalten vernachlässigt werden. Zusätzlich spielen Veröffentlichungen in nationalen Sprachen eine wichtige Rolle bei der Verbreitung von Erkenntnissen in möglicherweise nur lokal aktiven oder relevanten Subdisziplinen und bei der Beschreibung von Konzepten, die sprachabhängig sind (z.B. Emotionen).

Nicht nur die Sichtung, auch bereits das Auffinden benötigter Informationen stellt ein großes Hindernis dar, da Nutzer in der Regel nur solche Dokumente finden, die Begriffe enthalten, nach denen sie auch gesucht haben. Dokumente in Sprachen, die ein Nutzer nicht spricht, bleiben so in aller Regel für diesen unsichtbar.

Um diesem Umstand zu begegnen, hat das Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) zusammen mit weiteren Partnern im Jahr 2013 die vertikale Suchmaschine *PubPsych* (http://www.pubpsych.eu) entwickelt. PubPsych dient als Open-Access-Portal zur Bereitstellung von psychologischer Literatur, Tests, Behandlungsprogrammen und Forschungsdaten. Das Portal beinhaltet keine Volltexte, sondern die bibliographischen Metadaten zu einzelnen Dokumenten, die von mehreren Datenlieferanten bereitgestellt und im Index von PubPsych aggregiert werden. Die beteiligten Partner sind das *Institut de l'Information Scientifique et Technique* (INIST-CNRS) in Frankreich, das *Centro de Ciencias Humanas y Sociales* (CCHS-CSIC) in Spanien, die Norwegische Nationalbibliothek, die *Data Archiving and Networked Services* (DANS) in den Niederlanden, die *U.S. National Library of Medicine* (NLM) sowie das *Education Resources Information Center* (ERIC) aus den USA.

Jeder bibliographische Datensatz in PubPsych ist in mindestens einer der Sprachen Deutsch, Englisch, Französisch oder Spanisch beschrieben. Insgesamt beinhaltet PubPsych Verweise auf Dokumente in über 50 Sprachen. Zusätzlich zu diesem multilingualen Inhalt bietet PubPsych zwei weitere Funktionalitäten an, um Mehrsprachigkeit zu unterstützen: (a) die Weboberfläche ist sowohl in Englisch, Deutsch, Französisch als auch in Spanisch verfügbar; (b) manche nicht-englischen Einträge beinhalten zusätzlich zu den Originaldaten einen englischen Titel und englische Verschlagwortungen. Da angenommen wird, dass Englisch die *lingua franca* in der wissenschaftlichen Kommunikation ist, soll durch Funktion (b) ermöglicht werden, dass eine Suche nach englischen Begriffen die umfangreichste Trefferliste liefert. Nichtsdestotrotz ist ein großer Anteil der Datensätze in PubPsych nur über eine Suche in der Sprache des jeweiligen Datensatzes erreichbar.

Die zum Zeitpunkt des Projektbeginns (2016) aktuelle Implementierung von PubPsych unterstützte kaum den multi- bzw. cross-lingualen Zugriff auf die Daten, wodurch das eigentliche Potenzial dieses Portals nicht ausgeschöpft werden konnte. Insbesondere das cross-linguale Mapping von kontrolliertem Vokabular war nicht möglich. Bedingt durch diese Sprachbarriere

konnten mit einer Suchanfrage (z.B. in Deutsch) oftmals nicht alle relevanten Dokumente gefunden werden, wenn diese nur in einer anderen Sprache vorlagen.

Das beantragte Forschungsprojekt „Overcoming language barriers – Cross-lingual search of bibliographic metadata“ zielte darauf ab, eine Lösung für die Nutzung von multilingualen Datenbeständen aus dem Bereich der Psychologie zu finden. Hierfür sollten vier verschiedene Ansätze des cross-lingualen Information Retrievals (CLIR) entwickelt, prototypisch implementiert und evaluiert werden. Die um diese Ansätze erweiterte Suchmaschinentechnik sollte nicht nur die theoretische Leistungsfähigkeit (u.a. Precision und Recall) des IR-Systems verbessern, sondern gleichzeitig auch aus Nutzersicht einen Mehrwert bieten. Das Projekt war daher empirisch ausgelegt; im Mittelpunkt stand die kontrollierte, systematische Untersuchung von verschiedenen Ansätzen zur Verbesserung eines produktiven Systems, das frei verfügbar genutzt werden kann. Zusätzlich sollte das Projekt die Nutzbarkeit der gewonnenen Erkenntnisse in anderen, ähnlichen Portalen darstellen.

Konzept und Ansätze

Die Implementierung von CLIR erlaubt es Nutzern von PubPsych ihre Informationsbedürfnisse in der von ihnen bevorzugten Sprache zu formulieren, jedoch ebenfalls solche Dokumente zu finden, die nicht in der Sprache der Suchanfrage vorliegen.

Eine Möglichkeit, CLIR umzusetzen, besteht in der Verwendung von Verfahren der maschinellen Übersetzung. Dabei können zwei grundlegende Vorgehensweisen unterschieden werden: die Übersetzung des gesamten Datenbestands in alle relevanten Sprachen oder die Übersetzung der Suchanfrage. Ein weiterer Ansatz ist die Nutzung von hochgradig präzisen, wissensbasierten Vokabularen, in denen Terme und Konzepte mehrsprachig hinterlegt sind. Eine Suchanfrage wird dann nicht maschinell übersetzt, sondern im Vokabular nachgeschlagen und in die dort angegebenen Zielsprachen übersetzt.

Obwohl die Nutzung von automatischen Übersetzungsverfahren auf Grund von Adäquatheitsproblemen der Übersetzungen lange Zeit als nicht erfolgversprechend angesehen wurde, ließen aktuelle Entwicklungen im Bereich der statistischen maschinellen Übersetzung (SMT) diesen Ansatz zu Projektbeginn wieder als untersuchungswürdig erscheinen. Das beantragte Forschungsvorhaben schlug daher vor, aktuelle Verfahren der maschinellen Übersetzung und den wissensbasierten Ansatz zu kombinieren, mehrere Methoden der CLIR-Implementierung zu testen und zu evaluieren und somit den am besten für PubPsych geeigneten Ansatz zu ermitteln.

Die folgenden vier Ansätze können grundsätzlich unterschieden werden. Anpassungen und Kombinationen sind denkbar und wurden im Laufe des Projektes entwickelt:

(1) Übersetzung der Suchanfrage: Sobald eine Suchanfrage abgesendet wird, erweitert das Portal diese Anfrage durch Hinzufügen aller passenden Übersetzungen der Terme oder Phrasen, die in der Anfrage enthalten sind. Diese angepasste Version der Suchanfrage wird dann an den Index weitergeleitet. Fehler in der Übersetzung sind an dieser Stelle nachteilig, da sie vom Nutzer nicht gewünschte Begriffe in die Anfrage einführen.
(2) Übersetzung der Metadaten: Ein weiterer Ansatz übersetzt nicht die Suchanfrage, sondern den gesamten relevanten Datenbestand (z.B. Titel, Abstracts, Schlagwörter) des Suchmaschinenindex. Dadurch werden die Dokumente um Metadaten angereichert, die nicht in der Originalsprache der Dokumente stehen, und die Sprache der Suchanfrage ist nicht mehr von Relevanz. Bei vier Sprachen (d.h. 12 Sprachpaaren)

bedeutet dies eine Übersetzung in 12 Richtungen. Die Geschwindigkeit des Information Retrieval wird bei diesem Ansatz kaum beeinträchtigt, da keine Online-Übersetzung der Suchanfrage stattfinden muss. Die Qualität einzelner Übersetzungen hängt jedoch von der Verfügbarkeit von Sprachpaar-Daten ab, mit denen die SMT-Systeme trainiert werden; Sprachpaare mit wenig parallelen Daten werden in schlechterer Qualität übersetzt.

(3) Englisch als Brückensprache: Bei diesem Ansatz werden sowohl die Suchanfrage als auch der Inhalt des Index ins Englische übersetzt. Verglichen mit Ansatz 2 sind hier nur drei Übersetzungsrichtungen und somit -systeme notwendig (En-De, En-Fr, En-Es). Für diesen Ansatz ist mit der größten Verfügbarkeit von Trainingsdaten zu rechnen, da in jedem Sprachpaar Englisch involviert ist. Dadurch ist die Übersetzung ins Englische möglicherweise am besten. Außerdem ist rund die Hälfte der Metadaten in PubPsych bereits in Englisch verfügbar. Da sowohl die Suchanfrage als auch die Metadaten übersetzt werden, können Fehler jedoch auf beiden Seiten auftreten.

(4) Wissensbasierter Ansatz: In diesem Ansatz wird ein multilingualer Thesaurus bzw. ein multilinguales Lexikon verwendet. Dort ist das in den PubPsych-Dokumenten enthaltene kontrollierte Vokabular in den vier Projektsprachen enthalten; jeder dieser Begriffe wird mit seinen Übersetzungen in den drei anderen Sprachen aufgeführt. Bei der Implementierung dieses Ansatzes muss eine Suchanfrage zunächst auf die in dem Lexikon enthaltenen Begriffe abgebildet werden. Ein Vorteil dieses Ansatzes ist, dass die Suchanfrage weiterhin in der vom Nutzer bevorzugten Sprache formuliert werden kann und dass Begriffe verwendet werden, die auch tatsächlich in den Dokumenten enthalten sind. Da die Datensätze in PubPsych jedoch mehr als ein kontrolliertes Vokabular nutzen, müsste zunächst eine Konkordanz zwischen diesen Vokabularen erstellt werden, um ein möglichst einheitliches und umfassendes Lexikon zu erhalten. Der wissensbasierte Ansatz kann mit den anderen drei Ansätzen kombiniert werden.

In diesem Projekt wurde, nach unserem Wissen zum ersten Mal, die Anwendung aller vier Ansätze untersucht, um eine möglichst optimale Implementierung cross-lingualer Funktionalitäten in einer Suchmaschine für wissenschaftliche Veröffentlichungen zu erreichen. Das zum Zeitpunkt des Projektbeginns verwendete PubPsych-System diente dabei als Basis für das zu entwickelnde System.

1. Entwicklung der durchgeführten Arbeiten, einschließlich Abweichungen von der ursprünglichen Planung

Bevor ein System zur automatischen Übersetzung der Metadaten entwickelt werden konnte, musste der Datenbestand in PubPsych zunächst analysiert, aktualisiert und für die Zwecke des Projekts angepasst.

1.1. Repräsentation multilingualer Metadaten in PubPsych

Da es sich bei den in PubPsych enthaltenen Metadaten um einen fortlaufend aktualisierten und wachsenden Bestand handelt, musste zu Beginn des Projektes zunächst der momentane Status des Systems als Datenabzug exportiert und „eingefroren“ werden. Hierdurch wurde gewährleistet, dass sich am untersuchten Datenbestand während der gesamten Projektlaufzeit und insbesondere zwischen verschiedenen Untersuchungszeitpunkten der verschiedenen Übersetzungsansätze keine nicht-projektbedingten Änderungen ergaben. Gleichzeitig wurde die

dem PubPsych-System zugrundeliegende Software aktualisiert und für eine leichtere Bearbeitung angepasst. Der Grunddatenbestand beinhaltete zum Zeitpunkt des Datenabzugs 958.726 Datensätze.

Basierend auf diesen Daten wurden relevante Charakteristika des PubPsych-Systems ermittelt, die zur weiteren Analyse und Aufbereitung der Daten im Rahmen des Projektes notwendig waren. Hierzu zählten neben den allgemeinen, das System beschreibenden Daten (Anzahl der Datensätze, Datenlieferanten, vorhandene Metadatenfelder etc.) vor allem Angaben über Umfang und Art bereits vorhandener Multilingualität: Einige Datenlieferanten stellen in den an PubPsych übermittelten Metadaten bereits mehrsprachige Informationen – z.B. übersetzte Titel und/oder Abstracts – zur Verfügung, andere beschränken sich lediglich auf die Angabe der Originaldaten. Die Bandbreite reicht von sehr ausführlichen Angaben bis hin zu Datensätzen, bei denen keinerlei Sprachinformationen über den beschriebenen Datensatz zu finden sind.

Durch diese erste Analyse konnten bereits elementare Notwendigkeiten für eine zielgerichtete und zuverlässige automatische Übersetzung identifiziert werden. Ebenso ergaben sich Erkenntnisse zu einer uneinheitlichen Nutzung der Metadatenfelder beim Import der aus unterschiedlichen Quellen stammenden Daten. Da die Metadatenfelder zu Sprachangaben von Titeln, Abstracts oder dem den Metadaten zugrundeliegenden Volltext nicht bei allen Datensätzen ausgefüllt waren oder diese Angaben sich in unterschiedlichen Metadatenfeldern befanden, musste in einem ersten Schritt eine Vereinheitlichung der Sprachangaben stattfinden. Hierdurch wurde jedem Datensatz eine Originalsprache zugeordnet, die es nun ermöglichte, einen genaueren Überblick über die einzelnen in PubPsych-Einträgen vorhandenen Sprachen und ihre Verteilung zu erhalten.

Die bereits vorhandenen mehrsprachigen Daten wurden beibehalten, ihre Repräsentation im Index von PubPsych musste jedoch vereinheitlicht werden. Zu diesem Zweck wurden neue Metadatenfelder implementiert, die zur Aufnahme multilingualer Informationen dienen. An dieser Stelle wurde auch der Tatsache, dass zu einem späteren Zeitpunkt automatisch übersetzte Inhalte Eingang in den Suchmaschinenindex finden werden, Rechnung getragen: Beispiele für solche (neu eingeführte) relevante Metadatenfelder sind *TI_E* für einen nicht automatisch übersetzten englischen Titel oder *TI_E_from_S* für einen automatisch aus dem Spanischen ins Englische übersetzten Titel. Das Feld *TI_orig* beinhaltet hingegen den Originaltitel eines Dokuments, unabhängig von seiner Sprache. Diese neu hinzugefügten Felder erlaubten nicht nur die Abbildung unterschiedlicher Varianten der (automatischen) Übersetzung (z.B. unterschiedliche Übersetzungsrichtungen), durch sie wurde auch die Evaluation der einzelnen Ansätze des Projektes erleichtert: Indem einzelne Felder gezielt von der Indexierung oder der Suchabfrage ausgeschlossen werden konnten, ließen sich unterschiedliche Kombinationen der Ansätze auf ihre jeweilige Wirksamkeit testen. Die bereits in den PubPsych-Datensätzen enthaltenen mehrsprachigen Informationen wurden in die verbesserten Feldstrukturen überführt, um eine einheitliche Repräsentation der Daten zu gewährleisten.

1.2. Erstellung eines Goldstandards und Analyse realer Suchanfragen

Um ein System zur maschinellen Übersetzung von Texten evaluieren zu können, wird ein Goldstandard benötigt, der die von menschlichen Übersetzern erreichbare Übersetzungsqualität widerspiegelt. Aus dem Gesamtdatenbestand wurden daher diejenigen Datensätze extrahiert, die lediglich ein englisches Abstract und keine parallelen Übersetzungen beinhalteten. Aus diesen 448.732 Datensätzen wurde wiederum ein zufälliges Sample von 800 Abstracts erstellt, welches das relative Verhältnis der einzelnen Datenquellen zueinander berücksichtigte.

Diese 800 englischen Abstracts wurden bereinigt, auf Fehler kontrolliert, und dann von jeweils zwei Personen auf Satzebene in die anderen drei Projektsprachen Deutsch, Spanisch und Französisch übersetzt. Sie bilden den Goldstandard für die Evaluation maschineller Übersetzungen. Insgesamt handelt es sich um 7.195 Sätze mit 145.538 Wörtern. Zur Unterstützung der Übersetzerinnen wurde eine Webseite zur Suche nach Übersetzungen von Fachbegriffen im MeSH Thesaurus und im APA Thesaurus erstellt.

Um das Nutzerverhalten bei der Nutzung von PubPsych besser zu verstehen und auf Multilingualität hin zu untersuchen, wurden neben den Eigenschaften des Datenbestandes auch die Charakteristika realer Suchanfragen an das System untersucht. Gleichzeitig konnte auf diese Weise der Ansatz der Online-Übersetzung näher definiert und methodisch eingegrenzt werden.

Die Auswertung der PubPsych-Query-Logs zwischen dem 01. Januar 2014 und dem 31. Dezember 2016 zeigte insgesamt 553.799 Anfragen in 154.495 Sessions. 378.500 dieser Anfragen waren unterschiedlich. In einer zufälligen Auswahl von 500 Anfragen wurde für jede einzelne Anfrage durch zwei Personen unabhängig voneinander die Sprache bestimmt. Unklare oder mehrdeutige Sprachen wurden ebenfalls markiert. Des Weiteren wurden die Anfragen in die drei inhaltlichen Kategorien *informational*, *navigational* und *transactional* klassifiziert, um typische Verwendungsszenarien von PubPsych zu bestimmen.

Um auch für die Übersetzung von Suchanfragen einen Goldstandard erstellen zu können, wurde ein Sample von 261 englischen Anfragen bestimmt, die – äquivalent zu den zuvor beschriebenen Abstracts – ebenfalls von jeweils zwei Personen in die Sprachen Deutsch, Spanisch und Französisch übersetzt wurden. Eine Auswahl von 50 dieser Anfragen wurde zur Evaluation der Systeme genutzt (siehe Abschnitt 2.7). Für diese 50 Anfragen wurden daher Kurzbeschreibungen des durch sie repräsentierten Informationsbedürfnisses (sogenannte *topic descriptions*) angefertigt,

1.3. Generierung eines viersprachigen Lexikons

Sowohl der Ansatz zur Übersetzung der Inhalte von PubPsych, als auch der Ansatz zur Übersetzung der Suchanfragen benötigten ein multilinguales Lexikon, das die Übersetzung bestimmter Begriffe in die vier im Projekt betrachteten Sprachen ermöglichte. Es bot sich an, bereits vorhandene Thesauri nachzunutzen, da diese oftmals in bereits übersetzten Formen vorliegen. Eine Komplikation bestand darin, dass die Datensätze in PubPsych mehr als nur ein Vokabular zur Beschreibung der Metadaten nutzen. Daher wurde für das Projekt ein viersprachiges Lexikon erstellt, das sich unterschiedlicher Quellen bediente und sowohl bei der Übersetzung von Inhalten als auch zur Übersetzung einer Suchanfrage verwendet werden konnte.

Grundlage dieses Lexikons war der *Medical Subject Headings Thesaurus* (MeSH) der US-amerikanischen *National Library of Medicine*. Dieser Thesaurus lag bereits in allen vier Sprachversionen vor und musste lediglich zu einer einzigen Datei zusammengeführt werden. Hierfür wurde die Software *MeSHMerger* entwickelt und im Rahmen des CLUBS-Projekts veröffentlicht. Eine weitere Quelle war der *Thesaurus of Psychological Index Terms* der *American Psychological Association* (APA Thesaurus). Einzelne PubPsych-Datenlieferanten nutzen angepasste bzw. übersetzte Varianten dieser beiden Thesauri oder selbst erstellte

kontrollierte Vokabulare. Soweit möglich, wurden deren Inhalte mit dem MeSH und APA Thesaurus aligniert.

Da insbesondere bei den Suchanfragen der Nutzer auch Begriffe auftauchen können, die nicht spezifisch für die Psychologie sind, muss ein nützliches Lexikon auch Konzepte aus anderen Bereichen beinhalten. Aus diesem Grund wurden weitere Daten in das viersprachige Lexikon integriert, wie z.B. Titel von Wikipedia-Artikeln aus dem Bereich Psychologie, Namen von Kategorien aus der Wikipedia und öffentlich zugängliche zweisprachige Lexika. Zuletzt wurden häufig in PubPsych vorkommende Terme, die nicht bereits in einem der o.g. Vokabulare enthalten waren, mithilfe des frei zugänglichen Übersetzungstool *DeepL*[1] übersetzt und dem viersprachigen Lexikon manuell hinzugefügt. Um die beste Kombination der in diesem Absatz genannten Ressourcen (sowohl domänen-spezifisch als auch domänen-unabhängig) ermitteln zu können, wurden unterschiedliche Zusammenstellungen evaluiert.

1.4. Mapping von Suchanfragen und Vokabularen

Suchanfragen sind in der Regel ungrammatische Reihungen von Wörtern und nutzen nicht nur diejenigen Begriffe, die in einem kontrollierten Vokabular enthalten sind. Daher musste ein Workflow entwickelt werden, der die Übersetzung einer Suchanfrage mithilfe des viersprachigen Lexikons ermöglichte.

Da bei kurzen Suchanfragen die Sprache der Anfrage nicht immer eindeutig zu bestimmen ist, ist das viersprachige Lexikon wie folgt aufgebaut: Zu jedem Begriff gibt es einen Eintrag, dem die Übersetzungen dieses Begriffes in die anderen drei Sprachen zugeordnet sind. Die Sprache des Begriffes selbst ist jedoch nicht definiert. Auf diese Art kann verhindert werden, dass zunächst die Sprache der Anfrage bestimmt werden muss. Stattdessen wird im Lexikon lediglich nachgeschlagen, ob für einen bestimmten Begriff Übersetzungen existieren. Sollte ein Begriff in mehreren Sprachen identisch sein, wird für den zu übersetzenden Begriff diejenige Sprache angenommen, die laut Analyse der Daten (s.o.) häufiger in PubPsych vorkommt.

Der Ablauf bei der Übersetzung einer Suchanfrage ist somit folgender: Zunächst wird im Lexikon überprüft, ob für die gesamte Einheit der Anfrage eine Übersetzung existiert. Ist dies nicht der Fall, werden einzelne Wörter aus der Anfrage extrahiert und für diese wiederum jeweils nach einer passenden Übersetzung im Lexikon gesucht. Wenn auch dieser Schritt nicht erfolgreich ist, wird das Wort (falls möglich) in den Singular konvertiert und erneut im Lexikon nachgeschlagen. Sollte auch das nicht erfolgreich sein, existiert keine Übersetzung für diesen Begriff und er wird nicht übersetzt. Dies ist z.B. der Fall bei Angaben wie ISBN, Zeitschriftentiteln oder Eigennamen.

Als Ergebnis dieses Prozesses erhält das System eine Suchanfrage, die neben den originalen Begriffen um die im Lexikon enthaltenen Übersetzungen angereicht wurde. Diese erweiterte Anfrage wird dann an den Suchmaschinenindex weitergeleitet und die passenden Dokumente werden zurückgegeben. Ein Vorteil dieses Workflows ist die Sprachunabhängigkeit, da Begriffe nicht anhand ihrer Sprache, sondern nur aufgrund ihres Vorkommens im Lexikon übersetzt werden. Die Abdeckung des viersprachigen Lexikons sowie die Qualität der mit dessen Hilfe durchgeführten Übersetzungen wurde evaluiert.

1.5. Korpora für die maschinelle Übersetzung

[1] https://www.deepl.com/translator

Zur Implementierung der Ansätze (2) und (3) wurden Verfahren der datenbasierten maschinellen Übersetzung benötigt, um Metadaten einer Ausgangssprache in eine andere Sprache zu übersetzen. Während des Projektes wurden sowohl diese Verfahren als auch der Effekt maschineller Übersetzungen auf die Retrievalqualität evaluiert (siehe Abschnitt 2.7).

Datenbasierte maschinelle Übersetzung basiert auf großen Menge paralleler und monolingualer Korpora, anhand derer die Übersetzungsmodelle erlernt werden können. Daher war die Zusammenstellung dieser Korpora ein zentrales Element des Projektes. Berücksichtigt wurden nicht nur Korpora aus dem Feld der Psychologie, sondern auch andere Domänen (z.B. Politik). Korpora, die nicht speziell den Bereich der Psychologie abdecken, sind einfacher zu erhalten und insbesondere in denjenigen Fällen hilfreich, in denen nur sehr wenige domänen-spezifische Daten verfügbar sind (z.B. für das Sprachpaar Englisch-Französisch).

Ein paralleles Korpus wurde aus den bereits in PubPsych enthaltenen multilingualen Metadaten erstellt. Es wurden nur die Titel und Abstracts der Dokumente berücksichtigt. Lagen zu einem Dokument Titel und/oder Abstract in mehr als einer Sprache vor, wurden diese parallelen Daten extrahiert und für das Training der maschinellen Übersetzung aufbereitet. Eine Besonderheit war die Aufteilung der Titel in zwei Metadatenfelder (Haupt- und Untertitel) in manchen Dokumenten, die nicht zwangsläufig für alle Sprachen dieses Dokumentes identisch war. Dies musste zunächst vereinheitlicht werden. Die Abstracts wurden auf Satzebene aligniert.

Der Umfang paralleler Daten in PubPsych ist in Tabelle 1 dargestellt. Kein Dokument lag in allen vier betrachteten Sprachen vor.

Tabelle 1: Multilingualität von Titeln und Abstracts in PubPsych

	En-De	**En-Es**	**En-Fr**	**De-Es**	**De-Fr**	**Es-Fr**	**En-Es-Fr**	**De-En-Fr**
Titel	307.370	25.680	45.324	7	50	2	2	6
Abstracts	47.218	16.934	189	0	0	105	105	0

Zusätzliche monolinguale Korpora wurden auf die gleiche Weise aus allen in PubPsych verfügbaren Titeln und Abstracts in den vier Sprachen erstellt. Eine Alignierung von Haupt- und Untertiteln musste hier aufgrund der fehlenden Parallelität nicht stattfinden. Zuletzt wurden alle erzeugten Korpora zusammengeführt und die 800 bereits manuell zu übersetzenden Abstracts (s.o.) entfernt, da diese ein eigenes Korpus zur Evaluation bildeten. Das endgültige Korpus bestand somit aus 957.926 Dokumenten. Hiervon wurden weitere 1.500 Dokumente entnommen, um ein Test- und Entwicklungsset für das zu erstellende Modell zu erhalten. Die restlichen Dokumente dienten zum Training des Modells.

Alle Daten wurden vorverarbeitet (Normalisierung, Tokenisierung, Truecasing), um sie in einheitlicher Form im weiteren Verlauf nutzen zu können.

Neben den Daten aus PubPsych wurden weitere Korpora verwendet: das *EMEA Corpus*[2] und das *Scielo Corpus*[3] (beide aus dem Bereich Medizin, Biologie, Lebenswissenschaften), *Europarl*

[2] http://www.emea.europa.eu
[3] http://www.scielo.org

Corpus und *United Nations Corpus* (Politik), *Common Crawl Corpus* und Wikipedia-Artikel (allgemeine Webdokumente), sowie *JR-ACQUIS*, *News Commentary*, *MODEL Rapid Corpus*, *EUbookshop*.

Folgende Schwachpunkte konnten während der Zusammenstellung der Korpora identifiziert werden:

(a) Die verfügbaren Übersetzungsdaten waren für die betrachteten Sprachpaare sehr ungleich verteilt. Für manche Sprachpaare existierten kaum geeignete parallele Daten (z.B. Deutsch-Französisch), in diesen Fällen müssen entweder Daten aus anderen Bereichen als der Psychologie oder Englisch als Brückensprache verwendet werden. Verfahren des statistischen maschinellen Lernens, wie sie ursprünglich im Konzept des Forschungsvorhabens vorgesehen waren, führen mit solchen Pivot-Ansätzen jedoch nicht zu zufriedenstellenden Ergebnissen.
(b) Da es sich bei den Titeln der Dokumente um in der Regel sehr kurze Texte handelt, funktionieren Verfahren der maschinellen Übersetzung hier nicht immer ausreichend gut.

Um die in Punkt (a) beschriebene Problematik zu entschärfen, wurden in der letzten Version des Modells zusätzliche allgemeine Korpora integriert. Auch wurde die Vorverarbeitung erweitert und angepasst, um systematische Fehler, die während der Entwicklung beobachtet wurden, zu beheben. Hierzu zählten z.B. die Entfernung von nicht-lateinischen Schriften oder die Bereinigung von HTML-Entitäten.

1.6. System zur maschinellen Übersetzung

Das beschriebene Forschungsprojekt verwendete Verfahren der maschinellen Übersetzung zur Umsetzung des cross-lingualen Information Retrievals (CLIR). Ursprünglich war die Nutzung von statistischen Methoden der maschinellen Übersetzung (SMT) geplant, da diese zur Zeit des Projektantrags den aktuellen Stand der Forschung widerspiegelten.

Während des Projekts kamen jedoch zahlreiche neue Forschungen und Erkenntnisse im Bereich des Deep Learning auf, sodass sich die Methode der neuronalen maschinellen Übersetzung (NMT) im Jahr 2016 zum Standard entwickelt und die SMT an Bedeutung verloren hatte. Aus diesem Grund wurde vom ursprünglichen Konzept abgewichen und im Rahmen des Projektes statt eines SMT-Systems ein NMT-System entwickelt. Zunächst war geplant, beide Varianten zu verfolgen und ihren Effekt auf die Retrievalqualität zu evaluieren, um schließlich den vielversprechendsten Ansatz zu implementieren. Als der hiermit verbundene Aufwand und das absehbar klare Ergebnis deutlicher wurden, beschränkte sich die folgende Entwicklung auf das erfolgversprechendere NMT-System.

Ein wesentlicher Vorteil der NMT-Architektur ist, dass – im Gegensatz zu SMT – nicht für jedes benötigte Sprachpaar ein eigenes Modell erstellt werden muss, sondern dass alle Sprachpaare in einem einzigen neuronalen Modell abgebildet werden können. Hierdurch kann auch zwischen Sprachen übersetzt werden, die in den Trainingskorpora nicht oder nur spärlich vorhanden sind. Insbesondere dieser letzte Punkt war im Projekt von großer Bedeutung (siehe auch den vorherigen Abschnitt über die Korpora für die maschinelle Übersetzung). Ein Nebeneffekt ist außerdem, dass Übersetzungen durch NMT in der Regel einen flüssigeren und somit verständlicheren Eindruck machen. Dies könnte von Bedeutung sein, falls Ansatz 2 (Übersetzung der Metadaten im Index) die besten Retrievalergebnisse erzeugt und die

automatischen Übersetzungen den Nutzern von PubPsych angezeigt werden sollen. Ein wesentlicher Nachteil von NMT-Systemen ist die deutlich längere Zeit, die das Training benötigt.

1.7. Evaluation

Die Evaluation der vier Ansätze des Projektes erfolgte sowohl intrinsisch als auch extrinsisch. Die *intrinsische* Evaluation wurde verwendet, um die Qualität bzw. Leistungsfähigkeit der automatischen Übersetzung und des Mappings zwischen Suchanfragen und Vokabularen zu ermitteln. Die *extrinsische* Evaluation überprüfte den Effekt eines Ansatzes auf die Qualität des Information Retrievals in PubPsych. Hierbei kann die Evaluation sowohl mit Fokus auf das System selbst stattfinden (d.h., ob sich das System formal betrachtet verbessert bzw. ändert, z.B. durch Anzeige von mehr Suchergebnissen) als auch bezogen auf die Nutzer (d.h., ob die Nutzer subjektiv eine Verbesserung des Systems empfinden, z.B. durch Anzeige von relevanten, dem Nutzer zuvor unbekannten Dokumenten). Die Kombination der Ergebnisse beider Evaluationsarten lässt den Schluss zu, welches Verfahren das am besten geeignete für die Suchmaschine PubPsych ist.

Folgende Evaluationen wurden durchgeführt:

- Intrinsische Evaluationen (z.B. BLEU-Scores) der Systeme zur automatischen Übersetzung. Darüber hinaus wurden Übersetzungsfehler und –auffälligkeiten stichprobenweise manuell überprüft, um die Systeme schrittweise anzupassen und zu verbessern.
- Extrinsische Evaluationen der Verfahren, die für die vier einzelnen Ansätze entwickelt wurden.
- Bestimmung des jeweils besten Verfahrens für die vier Ansätze.
- Vergleich dieser „Gewinner-Verfahren“, um den erfolgreichsten Ansatz zu ermitteln.

Der erfolgreichste Ansatz wurde dadurch bestimmt, dass alle Gewinner-Systeme dieselben 50 Suchanfragen (jeweils in allen vier Projektsprachen) verarbeiten mussten und die so erhaltenen Dokumente von Nutzern auf ihre Relevanz für die jeweilige Suchanfrage beurteilt wurden. Es wurde eine dreistufige Likert-Skala verwendet (Highly relevant, partially relevant, non relevant). Zum Schluss konnte die Qualität der Treffermengen mit gängigen Maßen wie z.B. Precision und Recall beschrieben werden.

Grundlage für die Evaluationen des IR-Systems war eine Baseline, die das PubPsych-System zu Beginn des Projektes – also ohne cross-linguale Mechanismen – widerspiegelte. Die Trefferlisten dieser Baseline dienten als Maßstab zur Beurteilung der während des Projektes implementierten Ansätze.

Hervorzuheben ist, dass keine belastbare, manuelle Evaluation der Übersetzungsqualität der durch das beste NMT-System erzeugten Übersetzungen stattfand. Diese Systeme wurden ausschließlich durch gängige Scores evaluiert (mit den o.g. stichprobenartigen Ausnahmen), da das Forschungsvorhaben als Ziel die Optimierung des Information Retrievals und nicht der Übersetzungssysteme hatte.

Für Ansatz 1 (automatische Übersetzung der Suchanfragen mithilfe eines viersprachigen Lexikons) wurde zusätzlich die Qualität der Übersetzung der Anfragen evaluiert. Grundlage der intrinsischen Evaluationen war hier das Korpus mit den 261 manuell übersetzten Suchanfragen (siehe Abschnitt 2.2). Eine manuelle Beurteilung der Übersetzungsqualität durch Nutzer war zunächst angedacht, wurde dann aber verworfen, da der Fokus wie im vorherigen Absatz bereits beschrieben auf der Evaluation der vollständigen Systeme lag. Aus demselben Grund

wurde auch das Mapping der einzelnen Thesauri bzw. Vokabulare nicht separat evaluiert. Hier wurde lediglich die Abdeckung des viersprachigen Lexikons, d.h., der Anteil der Suchanfragen, deren Terme im Lexikon zu finden sind, ermittelt.

Da keine Software identifiziert werden konnte, die eine einfache Durchführung der für dieses Projekt benötigten Relevanzbewertungen erlaubte, musste zunächst ein neues Tool programmiert werden, das den Upload von Suchanfragen und Ergebnislisten und die einzelne Bewertung von Anfrage-Dokument-Paaren durch Nutzer ermöglichte. Hierdurch verschob sich der geplante Beginn der finalen Evaluationen um einige Wochen nach hinten.

Mit Hilfe eines Pilot-Experimentes wurden die Software, der Evaluationsprozess und die Auswertung geprüft. Hierdurch ergaben sich noch weitere Anpassungen, z.B. in Bezug auf eine einheitlichere und verständlichere Beschreibung der *topic descriptions*. Daher wurde erst nach Abschluss und Auswertung des Pilot-Experiments die finale Evaluation mithilfe der aus dem Fachgebiet der Psychologie stammenden Nutzerinnen begonnen.

Im Projektantrag war die (jährliche) Evaluation einzelner Zwischenstufen bzw. Prototypen des zu entwickelnden cross-lingualen PubPsych-Systems vorgesehen. Aufgrund des experimentellen Aufwands, den belastbare Relevanzbewertungen einnahmen, musste auf diese Zwischenevaluationen verzichtet werden. Es wurde mithin nur der finale Prototyp bezüglich seiner Retrievalqualität aufwändig evaluiert.

1.8. CLuBS Abschluss-Workshop

Die bis Anfang Juni 2019 gewonnenen Erkenntnisse und Ergebnisse des Projektes wurden der Fachcommunity am 07. Juni 2019 im Rahmen eines von den Projektpartnern organisierten und durchgeführten Workshops mit 35 Teilnehmenden am *Deutschen Forschungszentrum für Künstliche Intelligenz* (DFKI) in Saarbrücken vorgestellt.

2. Erreichte Ergebnisse und Diskussion im Hinblick auf den aktuellen Forschungsstand

2.1. Suchanfragen in PubPsych

Die Nutzung wissenschaftlicher Suchmaschinen unterscheidet sich deutlich von der Nutzung von digital verfügbaren Bibliothekskatalogen oder von Suchmaschinen, die allgemeine Web-Dokumente indexieren. Dies liegt in den unterschiedlichen Anforderungen der jeweiligen Nutzergruppen begründet. Wissenschaftler suchen nach anderen, häufig fachspezifischen Begriffen und können mit den Funktionalitäten wissenschaftlicher Suchsysteme (z.B. Suche in einzelnen Metadatenfeldern, Phrasensuche, boolesche Operatoren) umgehen. Auch in anderen Studien beobachtete Charakteristika der Nutzung wissenschaftlicher Suchmaschinen konnten für PubPsych gezeigt werden.

Die Auswertung der PubPsych-Logs des Zeitraums 01. Januar 2014 bis 31. Dezember 2016 ergab, dass in 154.495 Sessions 553.799 Suchanfragen (378.500 davon unterschiedlich) an das PubPsych-System gestellt wurden. Ein Median von 2 Anfragen wurde pro Session gestellt, die durchschnittliche Länge einer Anfrage betrug dabei 3,6 Tokens (einfache Suche) bzw. 4,9 Tokens (erweiterte Suche). Wie in mehreren Studien bei anderen akademischen Suchmaschinen beobachtet, konnte auch für PubPsych gezeigt werden, dass der Großteil der Anfragen der Kategorie *informational* zuzuordnen ist: 88.4 % der Suchen fielen in diese Gruppe.

2.2. Viersprachiges Lexikon, Übersetzung von Anfragen und kontrolliertem Vokabular

Neben den Einträgen des MeSH-Thesaurus wurden Titel und Kategorien aus der Wikipedia in das viersprachige Lexikon übernommen. Aus dem Bereich Psychologie und Gesundheit konnten 81.369 Titel, die in allen vier Projektsprachen vorlagen, extrahiert werden. Des Weiteren wurden 38.038 alignierte Kategorienamen hinzugefügt und 5.576.686 alignierte Einträge aus Wikidata in das Lexikon überführt.

Eine weitere Quelle für nicht domänen-spezifisches Vokabular waren Wörterbücher der Open-Source-Plattform *Apertium*[4] (insgesamt 25.593 zusätzliche Einträge), sowie innerhalb des Projekts manuell übersetzte Terme, die in den in PubPsych vorhandenen kontrollierten Vokabularen auftauchten (insgesamt 16.532 zusätzliche Einträge).

Tabelle 2 fasst die Abdeckung der in PubPsych vorhandenen Kategorien und Deskriptoren einzelner PubPsych-Datenquellen durch zwei Varianten des so erstellten viersprachigen Lexikons zusammen (*MeSH* = Verwendung des MeSH-Thesaurus; *QuadLex* = viersprachiges Lexikon, inkl. MeSH, Wikipedia, Wikidata, Apertium und manuellen Einträgen). Die zweite Spalte gibt im Lexikon enthaltene Deskriptoren und Klassen an, die dritte Spalte die im Lexikon enthaltenen einzelnen Tokens der Deskriptoren und Klassen.

Tabelle 2: Abdeckung des kontrollierten Vokabulars durch das viersprachige Lexikon

Datenquelle	**Deskriptoren und Klassen** (übersetzt)	**Tokens** (übersetzt)
MeSH – ACCNO	344.453 (30,4 %)	1.325.648 (70,9 %)
QuadLex – ACCNO	598.348 (52,9 %)	1.867.350 (99,8 %)
MeSH – DFK	544.275 (33,3 %)	2.043.618 (77,2 %)
QuadLex – DFK	917.681 (56,1 %)	2.645.782 (99,9 %)
MeSH – NORART	5.630 (24,6 %)	34.048 (86,9 %)
QuadLex – NORART	7.263 (31,8 %)	39.167 (100 %)
MeSH – PDID	197 (43,9 %)	623 (80,5 %)
QuadLex – PDID	287 (63,9 %)	774 (100 %)
MeSH – PMID	2.987.945 (86,9 %)	5.007.120 (97,1 %)
QuadLex – PMID	3.096.000 (90,0 %)	5.160.435 (100 %)

Es zeigte sich, dass mit dem kompletten viersprachigen Lexikon bei Verwendung einzelner Tokens bis zu 100 Prozent des kontrollierten Vokabulars übersetzt werden konnten. Dennoch ist zu beachten, dass mit dieser Methode nicht in jedem Fall korrekte Übersetzungen erreicht werden. Gründe hierfür können z.B. mehrdeutige Begriffe oder Fehler in den Vokabularen sein. Auch ist eine Verknüpfung der Übersetzungen zweier einzelner Tokens nicht zwangsläufig deckungsgleich mit der korrekten Übersetzung des kompletten Deskriptors.

Ein ähnliches Bild zeigte sich bei der Übersetzung von Suchanfragen mit Hilfe des viersprachigen Lexikons: Bei ausschließlicher Nutzung des MeSH-Thesaurus und ohne Trennung der Suchanfrage in einzelne Wörter konnten 7,7 % der Anfragen übersetzt werden;

[4] https://github.com/apertium

bei zusätzlicher Nutzung der nicht domänen-spezifischen Vokabulare erhöhte sich dieser Anteil auf 14,9 %. Wenn die Suchanfrage nicht als Ganzes im Lexikon nachgeschlagen wurde, sondern auch die Überprüfung einzelner Wörter erlaubt wurde, erhöhte sich der Anteil auf 85,0 %.

Um die Angemessenheit dieses Ansatzes zu prüfen, wurde die Übersetzungsqualität der mit dem Lexikon übersetzten Suchanfragen für 500 Anfragen manuell kontrolliert. Es ergab sich eine durchschnittliche *Adequacy* von 1,4 auf einer Skala von 0-2, d.h., dass in der Mehrzahl der Anfragen Suchterme korrekt übersetzt wurden.

Dieselben 500 Anfragen wurden manuell übersetzt, um eine intrinsische Evaluation unter Nutzung des BLEU-Scores zu ermöglichen. Der beste BLEU-Wert von 59,82 wurden für einzelne Tokens und bei Benutzung des viersprachigen Lexikons aus Tabelle 2 erreicht. Bemerkenswert an dieser Stelle ist, dass dieser Wert sogar die zum Vergleich herangezogenen BLEU-Scores der beiden Übersetzungssysteme von *Google Translate* (56,66) und *DeepL* (53,58) übersteigt.

2.3. Übersetzung der bibliographischen Metadaten

Für die Übersetzung der bibliographischen Daten (Titel und Abstracts) wurden mehrere Übersetzungsmodelle, sowohl klassisch statistische, wie auch neuronale, trainiert. Nach der automatischen Evaluation der Übersetzungsqualität mit Metriken wie BLEU, TER und METEOR wurde ein neuronales System mit einer transformer big-Architektur integriert Das transformer-Modell wurde 2017 während des laufenden Projektvorhabens entwickelt und hat sich nachfolgend im Natural Language Processing als leistungsfähigste Architektur für die meisten Aufgaben im sequence-to-sequence-Bereich, wie maschineller Übersetzung, etabliert.

Mit diesem System wurden state-of-the-art Ergebnisse für die Übersetzungspaare Englisch-Deutsch und Englisch-Spanisch erreicht, mit etwas niedrigerer Qualität für Englisch-Französisch im Falle von Abstractübersetzungen. Die Ursache für diesen Qualitätsabfall war der Mangel an parallelem, französischem Trainingsmaterial in PubPsych (vgl. Tabelle 1).

Es wurden drei Techniken benutzt, um die Übersetzungsqualität der neuronalen Modelle für den spezifischen Fall des Projekts zu maximieren. Die erste, **back-translation**, dient dazu den Effekt weniger Trainingsdaten zu verringern und gleichzeitig das Modell robuster gegen noise zu machen. Dafür wurde monolinguales PubPsych-Material mit dem besten dem Projekt verfügbaren allgemeinen Modell übersetzt und diese Übersetzungen dann zusammen mit dem verfügbaren multilingualen, parallelen Daten zum Training genutzt. Das finale Trainigskorpus ist daher eine Kombination aus den echt parallelen und den synthetischen Daten der back-translation. Für die Sprachpaare, die nicht Englisch beinhalten, sind regelmäßig die meisten Daten back-translations.

Als zweite Technik kam der Ansatz der **domain adaption** aus dem Bereich des transfer learnings dazu. Diese bedingt das Training eines allgemeinen, domänenunspezifischen Modells und folgend die iterative Anpassung mit domänenspezifischen Daten, die aus PubPsych extrahiert wurden. Mit diesen zusätzlichen Trainingsiterationen werden dann die Parameter des Modells für die Domäne optimiert.

Als letzte Technik wurde ein **Ensemble** aus verschiedenen Modellen (z.B. mit und ohne Domänenadaption und in verschiedenen Trainingszustände) genutzt, um eine robuste, endgültige Übersetzungssoftware zu erstellen. Das Ensemble wird im decoding genutzt, wo alle Modellparameter kombiniert werden um einen Satz zu übersetzen und die endgültige Übersetzung dann aus einer kombinierten *n*-Bestenliste gewählt wird.

Im Projekt wurde das Marian-Toolkit[5] genutzt. Das Projekt stellt das beste multilinguale Modell zur Übersetzung der in CLUBS abgedeckten Übersetzungspaare {en,de,fr,es}↔{en,de,fr,es} bereit. Die Übersetzungsqualität dieses Modells und der Vergleich mit kommerziell erhältlichen Systemen wie Deepl und Google Translate sind in der untenstehenden Übersicht dargestellt. Für das CluBS-System sind dort auch die Ergebnisse der automatischen Evaluation in drei Varianten für die sechs Übersetzungspaare

{en,de}↔{en,de}, {en,fr}↔{en,fr} und {en,es}↔{en,es} dargestellt. Dabei ist «P2» das allgemeine Modell trainiert mit back-translation, P2+DA enthält zusätzlich die Domänenanpassung und «P2 ens» bietet die Ergebnisse des Ensembles mehrerer Modelle.

	de2en				en2de			
Titles								
	BLEU	NIST	TER	MTR	BLEU	NIST	TER	MTR
Google	40.91	8.07	43.04	70.55	31.02	6.46	56.32	53.05
DeepL	40.86	8.04	43.57	70.43	30.38	6.34	57.17	52.52
P2	44.61	8.47	40.26	72.39	36.11	7.19	49.94	57.44
P2+DA	45.03	8.51	**39.95**	72.57	34.65	7.11	50.27	56.64
P2 ens	**45.84**	**8.55**	40.02	**72.80**	**36.44**	**7.28**	**49.10**	**58.20**
Abstracts								
	BLEU	NIST	TER	MTR	BLEU	NIST	TER	MTR
Google	18.74	5.20	79.25	45.44	13.32	4.19	86.07	32.46
DeepL	19.04	5.24	79.51	46.04	**14.23**	4.29	85.63	33.27
P2	18.62	5.32	75.46	46.01	13.55	4.37	81.20	33.55
P2+DA	18.52	5.31	74.94	46.07	13.75	4.39	80.52	33.78
P2 ens	**18.95**	**5.37**	**74.15**	**46.45**	14.10	**4.43**	**80.12**	**34.30**

	es2en				en2es			
Titles								
	BLEU	NIST	TER	MTR	BLEU	NIST	TER	MTR

[5] https://marian-nmt.github.io/

Google	42.96	8.02	41.73	73.23	49.92	8.45	39.20	69.97
DeepL	45.08	8.19	39.98	74.61	50.65	8.56	37.19	70.48
P2	46.89	8.40	38.60	74.86	53.44	9.02	33.85	72.62
P2+DA	47.36	8.48	37.92	75.24	53.98	9.03	33.28	72.73
P2 ens	**48.56**	**8.58**	**36.75**	**76.13**	**56.26**	**9.27**	**31.98**	**74.22**
Abstracts								
	BLEU	NIST	TER	MTR	BLEU	NIST	TER	MTR
Google	34.43	8.05	53.06	65.16	37.89	8.43	50.85	60.20
DeepL	**34.82**	8.10	52.82	**65.19**	**39.74**	8.63	49.56	61.34
P2	33.93	8.17	52.85	64.71	36.56	8.49	50.17	60.52
P2+DA	33.54	8.12	53.15	64.44	37.84	8.70	49.16	61.22
P2 ens	34.33	**8.22**	**52.65**	64.85	38.21	**8.74**	**48.79**	**61.75**

	fr2en				en2fr			
Titles								
	BLEU	NIST	TER	MTR	BLEU	NIST	TER	MTR
Google	47.33	7.19	39.70	75.50	42.64	7.08	42.95	65.98
DeepL	45.61	7.09	41.32	73.75	45.88	7.21	42.15	67.14
P2	45.72	7.11	40.90	74.96	43.43	7.06	43.11	66.11
P2+DA	46.37	7.22	40.43	74.41	44.74	7.16	41.55	66.93
P2 ens	47.17	7.34	39.20	74.97	45.35	7.24	41.21	67.32
Abstracts								
	BLEU	NIST	TER	MTR	BLEU	NIST	TER	MTR
Google	29.26	6.70	60.96	59.29	26.42	6.44	62.92	50.77
DeepL	**29.69**	**6.79**	**59.99**	**59.31**	**27.90**	**6.60**	**61.70**	**51.53**
P2	27.22	6.62	61.38	58.29	23.66	5.79	63.67	49.19
P2+DA	25.53	6.46	61.67	57.42	23.89	6.00	64.05	49.14
P2 ens	27.06	6.63	61.11	58.12	24.55	6.03	63.46	49.82

2.4. Evaluation des Gewinnersystems

Das finale Prototypensystem sollte als Ganzes einer menschlichen Evaluation unterworfen werden, um die reale Nützlichkeit des Ansatzes wissenschaftlich abzusichern. Das vorab durchgeführte Pilot-Experiment, das zur Überprüfung der Gestaltung des geplanten

Rating-Prozesses diente, lieferte mehrere Einsichten, die bei den endgültigen Relevanzbewertungen berücksichtigt wurden:

(a) Um einen Verzerrungseffekt durch falsche oder missverständliche maschinelle Übersetzungen auszuschließen, wurden den Bewerterinnen die Metadaten in der Originalsprache gezeigt. Dies bedeutet allerdings, dass für Deutsch, Spanisch und Französisch jeweils eigene Bewerterinnen benötigt wurden (es wurde davon ausgegangen, dass alle Bewerterinnen Dokumente in Englisch angemessen beurteilen können).
(b) Um die Relevanz eines Dokumentes möglichst umfassend beurteilen zu können, wurden so viele Informationen wie möglich benötigt. Daher wurde die Angabe von Schlagwörtern zur Präsentation der Dokumente hinzugefügt. Außerdem wurden englische Metadaten, die bei nicht-englischen Dokumenten bereits in den Originaldaten (d.h. nicht automatisch übersetzt) vorhanden waren, zusätzlich angezeigt.
(c) Es wurde angenommen, dass pro Person ca. 50 Dokumente pro Stunde bewertet werden können. (Dies stellte sich während der finalen Relevanzbewertungen jedoch als zu optimistisch heraus.)

Nach der entsprechenden Anpassung des Bewertungswerkzeugs wurde die endgültige Relevanzbewertung der durch die unterschiedlichen Systeme erhaltenen Dokumente gestartet. Tabelle 3 gibt einen Überblick über die Durchläufe und die implementierten Systeme. Zu beachten ist, dass nicht jeder der vier Ansätze in einem eigenen System repräsentiert wurde. Aussagen über die Qualität von Ansatz 3 (Englisch als Brückensprache) ließen sich beispielsweise durch die Kombination der einzelnen „MT Baseline 2“-Durchläufe gewinnen. Insgesamt wurden 3.904 Dokumente auf ihre Relevanz in Bezug auf eine Suchanfrage bewertet.

Tabelle 3: Übersicht über die evaluierten Systeme

System	Suchanfragen	Anmerkungen
Baseline System 1	je 50 in 4 Sprachen; manuell aus EN übersetzt	System zu Beginn des Projektes
Baseline System 2	je 50 in 4 Sprachen; manuell aus EN übersetzt	Baseline System 1 mit Anpassungen nach Beginn des Projektes
MT System Baseline 2 – DE	je 50 in EN, FR, ES; automatisch aus DE übersetzt	Baseline System 2, Übersetzungsrichtung DE-EN/FR/ES
MT System Baseline 2 – FR	je 50 in DE, EN, ES; automatisch aus FR übersetzt	Baseline System 2, Übersetzungsrichtung FR-DE/EN/ES
MT System Baseline 2 – ES	je 50 in DE, EN, FR; automatisch aus ES übersetzt	Baseline System 2, Übersetzungsrichtung ES-DE/EN/FR
MT System Baseline 2 – EN	je 50 in DE, FR, ES; automatisch aus EN übersetzt	Baseline System 2; Übersetzungsrichtung EN-DE/FR/ES

Content Translation System	je 50 in 4 Sprachen; manuell aus EN übersetzt	beinhaltete die Übersetzung aller Metadaten im Index
Query Translation System	je 50 in 4 Sprachen; manuell aus EN übersetzt	beinhaltete die Online-Suchanfragen-Übersetzung
Pilot (Baseline System 2)	50 Original-Anfragen in Englisch	beinhaltete insgesamt 100 Dokumente; diente zur Ermittlung des *Inter-Annotator-Agreements* (IAA) zwischen den Nutzerinnen

Nach Auswertung aller Relevanzbewertungen konnten gängige Maße des Information Retrieval berechnet werden, um die Leistungsfähigkeit der einzelnen Systeme miteinander zu vergleichen (siehe Tabelle 4).

Tabelle 4: IR-Maße der implementierten Systeme

System	R-Precision	Precision@10	Recall(10)	nDCG
Baseline System 1	0.660	0.660	0.660	0.569
Baseline System 2	0.651	0.651	0.651	0.567
MT System Baseline 2 – DE	0.637	0.637	0.637	0.565
MT System Baseline 2 – FR	0.570	0.570	0.570	0.497
MT System Baseline 2 – ES	0.566	0.566	0.566	0.554
MT System Baseline 2 – EN	0.519	0.519	0.519	0.504
Content Translation System	**0.830**	**0.830**	**0.830**	**0.702**
Query Translation System	0.714	0.714	0.714	0.606

Auf Grundlage dieser Maße wurden die einzelnen im Projekt untersuchten Ansätze miteinander verglichen. Als Ergebnis ließ sich festhalten, dass Ansatz 2 (Übersetzung der Metadaten) die relevantesten Ergebnisse lieferte.

2.5. Anmerkungen zur Evaluation

Relevanzbewertungen in einer spezialisierten Domäne wie der Psychologie sind komplex und zeitaufwändig. Dies liegt unter anderem daran, dass geeignete Nutzerinnen gefunden werden müssen, die das fachspezifische Vokabular verstehen und somit die Relevanz der Ergebnisse einer spezialisierten Suchmaschine überhaupt fachlich beurteilen können. Die Multilingualität in diesem Projekt verstärkt dieses Problem noch, da für jede Sprache eigene Nutzerinnen gefunden werden müssen (sofern nicht eine Person sämtliche im Projekt untersuchten Sprachen spricht).

Hinzu kommt, dass die aus einer realen Anwendung stammenden Daten nicht vollständig korrekt sind und sich somit erst im Prozess Fehler und Probleme zeigen, die vor der Relevanzbewertung behoben werden müssen. Beispielsweise führten falsche Angaben zur Sprachinformationen in Datensätzen dazu, dass diese Dokumente einer Bewerterin zur Relevanzbewertung vorgelegt wurden, welche die entsprechende Sprache nicht verstand. Solche Dokumente mussten identifiziert, separiert, korrigiert und dann der korrekten Bewerterin neu vorgelegt werden.

3. Mögliche Anwendungsperspektiven und denkbare Folgevorhaben sowie wirtschaftliche Verwertbarkeit der Ergebnisse

Nach Abschluss des Projektes wurden die Übersetzungen des erfolgreichsten Ansatzes in das PubPsych-Produktivsystem übernommen und somit über die Adresse http://www.pubpsych.eu weltweit verfügbar. Die aufwendig manuell erstellten Korpora (übersetzte Abstracts und übersetzte Suchanfragen in vier Sprachen) werden veröffentlicht und der Wissenschaft zur Nach- und Weiternutzung zur Verfügung gestellt.

Zu berücksichtigen ist, dass sich ein Angebot wie PubPsych beständig weiterentwickelt (sowohl inhaltlich als auch technisch) und dass die dort implementierten Funktionalitäten des CLIR ebenso regelmäßig für neue Anforderungen angepasst werden müssen. Beispielsweise könnte eine Kooperation mit neuen Datenlieferanten dazu führen, dass weitere Sprachen in das Portal integriert werden müssen. Das System zur automatischen Übersetzung müsste in diesem Fall um diese weiteren Daten erweitert werden, um weiterhin zuverlässige Übersetzungen produzieren zu können.

Die Überprüfung des Transfers der im Projekt gewonnenen Erkenntnisse auf andere, ähnlich gelagerte Anwendungsfälle (z.B. Suchportale anderer Fachdisziplinen) war bei Antragsstellung vorgesehen, konnte durch Verzögerungen zum Ende des Projektes jedoch nicht mehr im Detail untersucht werden. Die im Projekt entwickelten Prozesse wurden möglichst generisch angelegt, sodass davon auszugehen ist, dass eine Übertragung des hier dargestellten Verfahrens auf andere Systeme mit geringen Anpassungen (u.a. Berücksichtigung fachspezifischer Vokabulare bei der Entwicklung des Übersetzungsmodells) möglich ist. Eine Überprüfung dieser Annahme steht allerdings aus und wäre ein lohnendes Vorhaben. Insbesondere eine Übertragung der Erkenntnisse in Disziplinen, die dem Feld der Psychologie nahestehen (z.B. Medizin und Erziehungswissenschaften), könnte näher betrachtet werden. Hier dürften sich die größten Synergieeffekte und Nachnutzungsmöglichkeiten ergeben.

Suchportale existieren in ganz unterschiedlichen Anwendungskontexten. Informationssysteme für Bibliotheken weisen andere Erfordernisse auf als disziplinspezifische Angebote. Auch ist das Nutzerverhalten in diesen Systemen nicht zwangsläufig vergleichbar: In Bibliothekssystemen sind z.B. deutlich mehr Suchanfragen dem Typ *navigational* zuzuordnen, d.h., die Nutzer wissen bereits, was sie suchen, benötigen jedoch den Zugriff auf das entsprechende Dokument bzw. eine Auskunft zum Standort. Ob die Ansätze des Projektes auch in solchen veränderten Kontexten zielführend verwendet werden können, ist ein weiterer ungeklärter Punkt, der zukünftig näher untersucht werden könnte.

CLIR beinhaltet darüber hinaus viele weitere Faktoren, die im Rahmen des Projektes nicht verfolgt werden konnten, jedoch in möglichen Folgevorhaben untersucht werden könnten. Hierzu zählen z.B. die Berücksichtigung der vom Nutzer bevorzugten bzw. beherrschten

Sprachen, die Personalisierung des Suchportals, das Design des Suchinterfaces, die Darstellung der übersetzten Metadaten, die Anpassung des Suchalgorithmus oder die Bereitstellung von Möglichkeiten zum Nutzer-Feedback.

4. Beiträge von Kooperationspartnern

Die Arbeiten am Forschungsprojekt wurden von den Kooperationspartnern gemeinschaftlich mit unterschiedlichen Arbeitsanteilen und Verantwortlichkeiten durchgeführt.

Das Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin brachte sich mit Wissen zu domänenspezifischem und multilingualem Information Retrieval ein, plante die Evaluationen der einzelnen Ansätze und führte diese durch.

Der Lehrstuhl für Translationsorientierte Sprachtechnologie der Universität des Saarlandes steuerte die Expertise im Bereich der Computerlinguistik bei, übernahm die Konzeption, Entwicklung, Anpassung, Evaluation und Anwendung der Verfahren zur maschinellen Übersetzung und integrierte die Online-Übersetzung für Suchanfragen in PubPsych.

Das ZPID arbeitete in allen Arbeitspaketen und übernahm die Koordinierung des Projekts. Es stellte PubPsych als Anwendungsfall bereit, unterstützte die Implementierung der Ansätze in die Suchmaschine und koordinierte die Tätigkeiten durch Nutzer aus dem Fachgebiet der Psychologie (manuelle Übersetzungen und Relevanzbewertungen).

Durch Vorträge und Publikationen haben alle Kooperationspartner Ergebnisse des Projekts der wissenschaftlichen Community präsentiert.

Neben den Projektpartnern aus Trier, Berlin und Saarbrücken war an dem Vorhaben das Zentrum für Datenverarbeitung der Universität Mainz beteiligt, das, über das Land Rheinland-Pfalz, eine Seafile Cloud-Umgebung zum Austausch großer Dateien für das Projekt bereitstellte.

5. Qualifikationsarbeiten

Es wurden folgende projektbezogene Masterarbeiten erfolgreich abgeschlossen:

Ruiter, Dana (2019): Online Parallel Data Extraction with Neural Machine Translation. Master's Thesis, Saarland University, Saarbrücken.

Varga, Ádám Csaba (2017): Domain Adaptation for Multilingual Neural Machine Translation. Master's Thesis. Saarland University, Saarbrücken.

Yin, Jie (2017): Query log analysis of information search behavior in the psychological domain: a case study of PubPsych. Master's Thesis. Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin, Berlin.

6. Publikationen

España-Bonet, Cristina; Barrón-Cedeño, Alberto (2017): *Lump at SemEval-2017 Task 1: Towards an Interlingua Semantic Similarity*. Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), Vancouver, Canada, August 3rd-4th, 2017, pp. 144-149.

España-Bonet, Cristina; van Genabith, Josef (2017): *Going beyond zero-shot MT: combining phonological, morphological and semantic factors. The UdS-DFKI System at IWSLT 2017*. Proceedings of the 14th International Workshop on Spoken Language Translation, Tokyo, Japan, December 14th-15th, 2017, pp.15-22.

España-Bonet, Cristina; Varga, Ádám Csaba; Barrón-Cedeño; van Genabith, Josef (2017): *An Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification*. IEEE Journal of Selected Topics in Signal Processing 11(8), pp. 1349-1350. https://doi.org/10.1109/JSTSP.2017.2764273

España-Bonet, Cristina; Stiller, Juliane; Ramthun, Roland; van Genabith, Josef; Petras, Vivien (2018): *Query Translation for Cross-lingual Search in the Academic Search Engine PubPsych*. MTSR 2018, 12th International Conference on Metadata and Semantics Research, Limassol, Cyprus, October 23rd-26th, 2018, pp. 37-49. (Communications in Computer and Information Science, vol. 846). https://doi.org/10.23668/psycharchives.1062

Ruiter, Dana; España-Bonet, Cristina; van Genabith, Josef (2019): *Self-Supervised Neural Machine Translation.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 28th-August 2nd, 2019, pp.1828-1834.

Weichselgartner, Erich; Baier, Christiane; Ramthun, Roland (2017): *PubPsych: A Powerful Research Tool Providing Access to a Broad Supranational Body of Psychological Knowledge*. Datenbank-Spektrum 17(1), pp. 35-39. https://doi.org/10.1007/s13222-016-0244-3

7. Beiträge auf Tagungen und Kongressen

España-Bonet, Cristina; Barrón-Cedeño, Alberto (2017): *Lump at SemEval-2017 Task 1: Towards an Interlingua Semantic Similarity*. Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), Vancouver, Canada, August 3rd-4th, 2017, pp. 144-149.

España-Bonet, Cristina; van Genabith, Josef (2017): *Going beyond zero-shot MT: combining phonological, morphological and semantic factors. The UdS-DFKI System at IWSLT 2017*. Proceedings of the 14th International Workshop on Spoken Language Translation, Tokyo, Japan, December 14th-15th, 2017, pp.15-22.

Ruiter, Dana; España-Bonet, Cristina; van Genabith, Josef (2019): *Self-Supervised Neural Machine Translation.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 28th-August 2nd, 2019, pp.1828-1834.

Lüschow, Andreas; Ramthun, Roland (2019): *The CLuBS Use Case: PubPsych*. CLuBS Final Project Workshop, Saarbrücken, Germany, June 7th, 2019.

Stiller, Juliane; España-Bonet, Cristina (2019): *CLuBS. Cross-Lingual Bibliographic Search*. CLuBS Final Project Workshop, Saarbrücken, Germany, June 7th 2019.

Weichselgartner, Erich; Ramthun, Roland (2019): *Cross-lingual Search in the Psychology Search Engine PubPsych.* XVI. European Congress of Psychology, July 5th, Moscow, 2019

8. Software

MeSHMerger. A software to create a multilingual term list from the Medical Subject Headings (MeSH) and its translations. https://github.com/clubs-project/MeSHMerger

CLUBS Compa. A web application for comparing documents or search engine websites and for assessing retrieval quality of search engine results. https://github.com/alueschow/clubs-compa

9. Sicherung und Verfügbarmachung der Forschungsdaten

Die im Rahmen des Forschungsprojektes entstandenen Korpora (Übersetzungen von 261 Suchanfragen und 800 Abstracts) stehen unter https://doi.org/10.23668/psycharchives.2794 zur Verfügung.

Die domänenspezifischen Modelle zur automatischen Übersetzung der Titel und Abstracts stehen unter https://doi.org/10.5281/zenodo.3709164 zur Nachnutzung bereit.

Die im Rahmen des Projektes entstandene Projektdokumentation ist unter https://doi.org/10.23668/psycharchives.2746 verfügbar.

10. Pressemitteilungen und Medienberichte

CLUBS-Projektteam gewinnt Best-Paper-Award bei MTSR. leibniz-psychology.org, News vom 11. November 2018.

Workshop unterstreicht die Bedeutung cross-lingualer Informationsbeschaffung. leibniz-psychology.org, News vom 12. Juni 2019.