Eine Plattform zur Erschließung von wissenschaftlichen Inhalten mit Machine Learning und NLP: Der Kairntech “Sherpa”

Stefan Geißler
Kairntech – www.kairntech.com

Einführung

Der Sherpa – Leistungsfähige Lernverfahren – einfache Bedienung – vielseitige Einsatzszenarien

Maschinelle Lernverfahren (ML) setzen heute die Standards in vielen Bereichen der Informationsverarbeitung, so auch bei der Verarbeitung natürlicher Sprache (NLP): Thesaurus-basierte Indexierung, Dokumenten-Kategorisierung, die Extraktion von Personen, Orten oder Schlüsselbegriffen kann oft durch ML-Verfahren mit ansonsten unerreichter Qualität vorgenommen werden.

Gleichzeitig stellen ML-Verfahren oft hohe Anforderungen an die Verfügbarkeit von Beispielsdaten zum Training entsprechender Modelle sowie an die technische Kompetenz, um diese Modelle zu optimieren und produktiv einzusetzen.

Der **Kairntech Sherpa** stellt leistungsfähige und **vielseitige Lernalgorithmen** unter einer **leicht verständlichen Benutzeroberfläche** zur Verfügung und erlaubt Domänenexpert/innen das einfache und rasche Erstellen, Evaluieren sowie den Produktiveinsatz von NLP-Analysemodellen.

Die Analyse von wissenschaftlichen Aufsätzen, von Verträgen oder Rechnungen, von Nachrichtentexten, von Patenten oder Webinhalten sind nur einige der Einsatzzwecke, die mit dem Sherpa adressiert werden können.

Der Sherpa baut auf Verfahren unter anderem des “Deep Learning” auf, die erstklassige Ergebnisse auf öffentlich verfügbaren Evaluationsdaten wie Conll2003 erzielen. Einige der Verfahren, die Mitarbeiter aus dem Kairntech-Team implementiert haben, sind als open source in der Public Domain verfügbar – gleichzeitig sind sie als Teil des Sherpa ohne Programmierkenntnisse und eingebettet in eine moderne Weboberfläche zugänglich.

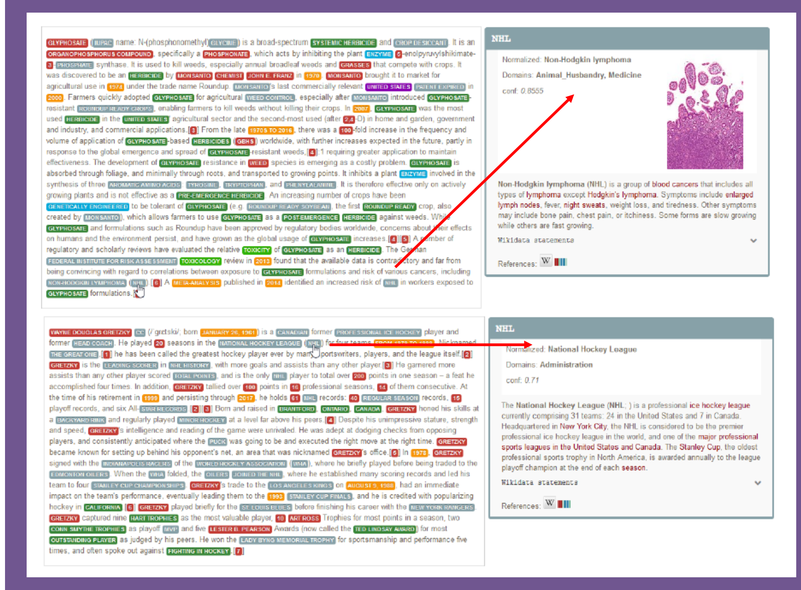
Der Sherpa kann “on premise” in der IT-Umgebung des Kunden installiert werden, oder in der Cloud und über ein umfangreiches API (Programmierschnittstelle) in Anwendungen integriert werden. Der Sherpa ist mehrbenutzer-fähig und unterstützt ein projekt-basiertes Arbeiten.

Wir stellen wichtige Komponenten und Einsatzszenarien des Sherpa in den folgenden Abschnitten vor.

Thesaurus-basiertes Indexieren

Extraktion, Normalisierung, Desambiguierung und Linking von Kernkonzepten

Die Erschließung von Inhalten mit Thesauri erfordert weit **mehr als nur die Suche nach Zeichenketten** im Text: Thesaurus-Terme können als Varianten (Samstag / Sonnabend) oder flektiert im Text (Haus / Häuser) vorkommen, ein und dieselbe Bezeichnung kann verschiedene Bedeutungen haben (Schloss: Vorrichtung zum Verschließen von Türen oder prachtvolles Gebäude) und schließlich profitiert die Erschließung von Inhalten, wenn die erkannten Begriffe mit Hintergrundinformation (z.B. Bild und CV einer prominenten Person, Geolokation eines Ortes, Webpage einer Firma, Strukturformel einer chemischen Substanz) verlinkt wird.



Der **Sherpa** erlaubt es Daten mit großen Thesauri – hier **Wikidata** mit fast 80 mio Konzepten in mehreren Sprachen – zu erschließen.

Im nebenstehenden Beispiel erkennt das System, dass die Zeichenkette “NHL” im ersten Fall für das “Non-Hodgkin Lymphoma” steht, im zweiten jedoch für die “National Hockey League”. Für diese Desambiguierung ist keinerlei manuelles Regelschreiben erforderlich. Das Lern-verfahren erkennt die richtige Lesart **anhand des Kontexts**.

Daten annotieren: Einfach und schnell

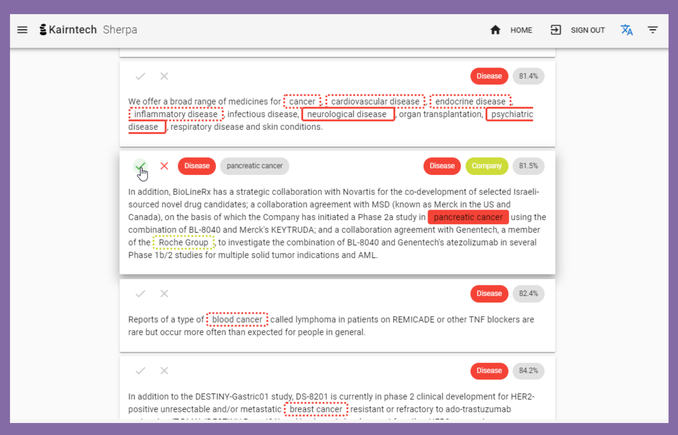
ML-Verfahren verlangen nach Trainingsdaten: Der Sherpa unterstützt Sie bei der Erstellung.

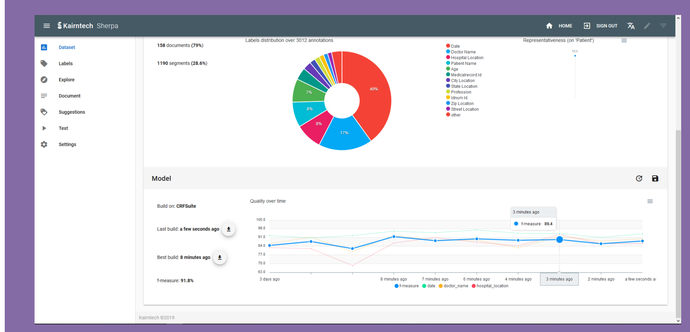
Moderne Lernverfahren sind oft vergleichsweise datenhungrig: Um das volle Potenzial zu nutzen, müssen sie oft an großen Mengen von Beispielen trainiert werden. Die manuelle Erstellung solcher Trainingsdaten kann überaus zeit- und kostenintensiv werden.

Die Sherpa-Benutzeroberfläche erlaubt es, mit einfachen Aktionen mit Maus und Keyboard rasch große Mengen von Daten zu annotieren.

Dabei lernt im Hintergrund ein Modell konstant mit und erzeugt auf Nachfrage neue Daten. Die Nutzer bekommen so stets Rückmeldung über die derzeit erreichte Qualität.

Durch den Einsatz von “**Active Learning**” wird die erforderliche Menge der nötigen Daten auf ein Minimum reduziert.



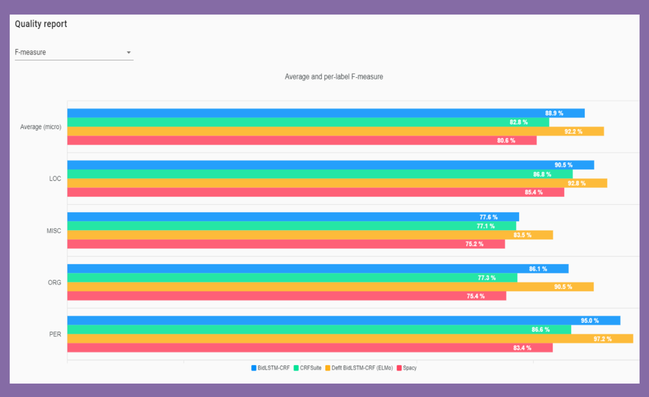


Der **Sherpa** vermittelt detaillierte Rückmeldung über die Corpusdaten, die Verteilung der unterschiedlichen Informationstypen und die Qualität der trainierten Modelle.

Nutzer können so zuverlässig erkennen, ob die zuletzt hinzugefügten Daten die Qualität weiter verbessert haben oder ob die selbstgesteckten Qualitätsziele bereits erreicht sind.

Unterschiedliche ML-Algorithmen weisen unterschiedliches Verhalten auf: Ein Ansatz kann sich womöglich sehr rasch trainieren lassen, erreicht aber letztlich nicht die gewünschte Qualität. Ein anderer dagegen benötigt sehr viel mehr Trainingszeit, liefert dann jedoch die besten Ergebnisse.

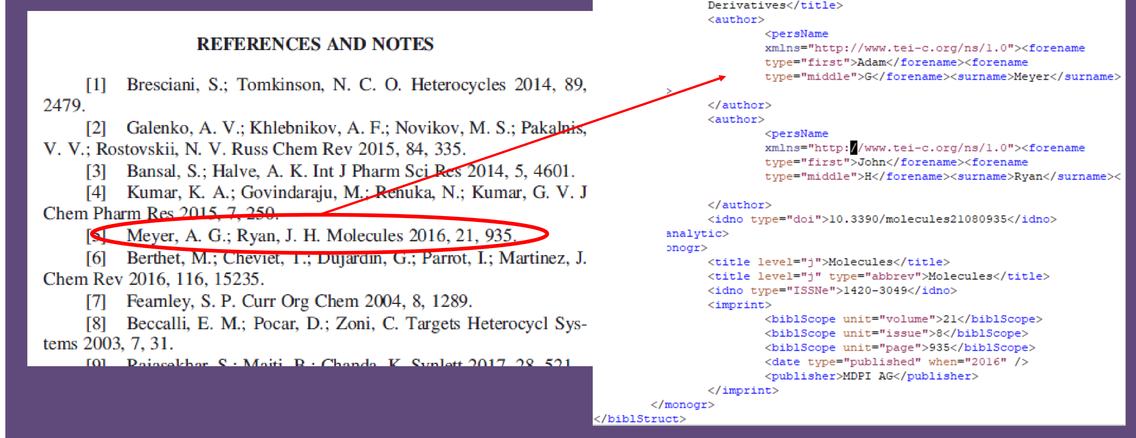
Der Sherpa erlaubt die Anwendung **unterschiedlicher Lernverfahren** auf den Daten eines Projekts und die Auswahl des für die jeweiligen Anforderungen besten Ergebnisses.



Der Sherpa kann wahlweise vorannotierte Daten importieren oder Rohdaten die erst durch die manuelle Bearbeitung zum Trainingscorpus werden. Die manuelle Annotierung ist dabei oft ein notwendiges Übel – zeitaufwändig aber leider erforderlich überall dort, wo Trainingsdaten nicht bereits verfügbar sind. Die Umsetzung des Annotationsprozesses im Sherpa legt daher großes Gewicht auf **intuitive und simple Benutzerführung**. Umfangreiche Tests mit Nutzern bestätigen die Benutzbarkeit: Nutzer melden uns: “Man bekommt regelrecht Lust, weitere Daten zu annotieren.”, “Besonders motivierend ist die **umgehende Rückmeldung** des Systems und die sich stetig verbessernden Resultate.”, “Regelrecht süchtig machend!”

Strukturerkennung von Dokumenten: PDF → XML

PDF ist der de-facto-Standard beim Austausch und der Speicherung von Dokumenten. Ein Nachteil des Formats ist jedoch, dass wichtige Strukturinformation zum Dokument, die bei der Erstellung noch vorlag, im PDF nicht mehr explizit ist, sondern erst wieder aufwändig erschlossen warden muss: Was ist der Titel des Dokuments? Wer sind die Autoren? Das Veröffentlichungsdatum? Gibt es ein Abstract? Welche anderen Werke werden zitiert und was waren wiederum deren Autoren?



Leistungsfähige ML-Verfahren verwenden hier eine Vielzahl von Informationen (Position im Dokument, Fonts, Layout, ...) um die Struktur der analysierten Dokumente zu ermitteln und das Dokument als TEI XML für weitere Auswertungen (z.B. Zitierungsnetzwerke) bereitzustellen. Im Beispiel oben wird das Literaturverzeichnis eines Aufsatzes analysiert: Das System erkennt die Autoren, den Titel und das Journal und kann automatisch den im PDF nur abgekürzt angegebenen Autor “Meyer, A. G.” als “Adam G. Meyer” expandieren.

Das eingesetzte System ist als open source verfügbar (https://github.com/kermitt2/grobid) und für den Einsatz auf wissenschaftlichen Dokumenten vortrainiert. Ein Einsatz auf anderen Dokumententypen ist möglich, aber erfordert das Anlernen der entsprechenden Modelle durch Kairntech.

Zum Einsatz kommen hier Kaskaden von ML-Modellen, die zunächst die Grobstruktur, dann die Details innerhalb der in Schritt 1 erkannten Abschnitte verarbeiten.

Kairntech – Wer wir sind.

Kairntech ist ein spezialisierter Anbieter von Software und Professionellen Dienstleistungen rund um die Themen Künstliche Intelligenz (KI) und NLP. Kairntech wurde im Dez 2018 von einem Team von Experten gegründet, die zuvor bereits über Jahre zusammen an vergleichbaren Themen gearbeitet hatten: Bei Xerox, IBM, INRIA und TEMIS.
Verschiedene Komponenten des Sherpa sind im Produktiveinsatz bei **CERN, NASA, ResearchGate, EPO, Mendeley**, ... Der Sherpa wird eingesetzt beispielsweise bei dem deutschen Pharmaunternehmen **Boehringer Ingelheim**.
Unser Logo – das Steinmännchen, engl. und frz. “Kairn” – spielt auf die Mission an, die wir uns gegeben haben: Wie ein Steinmännchen Bergsteigern den Weg durch unwegsames Gelände weist, so stehen wir bereit, unsere Kunden durch die manchmal unübersichtliche Welt der KI und NLP zu begleiten. Außerdem ist unser Hauptsitz in den Bergen in Grenoble, in Sichtweite des Mont Blanc.
Kontakt: www.kairntech.com, Email: info@kairntech.com