Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Informatik

Deep summarised and clustered Notice-to-airmen

Themenbereiche: Machine Learning, Natural Language Processing

Studierende: Cyrille Ulmi

Betreuungsperson: Daniel Pfäffli, MSE, HSLU, Rotkreuz

Experte: Ueli Amstutz, dipl. Ing. HTL, Opacc Software AG, Rothenburg

Auftraggebende: Hochschule Luzern - Informatik

Keywords: Text Summarization, Clustering, Sequence to Sequence, Machine learning, Natural Language Processing

1. Aufgabenstellung

NOTAM – Notice to airmen sind Meldungen, welche Akteure im Luftraum versenden, um auf mögliche Hindernisse aufmerksam zu machen. Linien-Pilote, sowie Freizeit-Pilote sind verpflichtet diese NOTAMs zu lesen.
NOTAMs werden in einer gekürzten Form geschrieben und beinhalten oft irrelevante Details, was zu einer Überforderung der Piloten führen kann.
Mittels neuronaler Netzwerke und der Natural Language Processing (NLP) Technik «Text Summarization» sollen die zuvor in englische Sprache übersetzten NOTAMs auf das Wesentliche gekürzt werden.
Anschliessend soll mittels «unsupervised» Clustering die zusammengefassten NOTAMs in sinnvolle Cluster übertragen werden. Schlussendlich wird für jeden dieser Cluster eine Zusammenfassung generiert. Das Resultat soll ansprechend als Grafik aufbereitet werden.

2. Ergebnisse

Im Rahmen der Bachelor Diplomarbeit wurden die folgenden Ergebnisse erarbeitet:

- Ein Datenkorpus der gesammelten NOTAMs wurde erstellt
- Ein Tool wurde erstellt, mit welchem die NOTAMs in eine ausgeschriebene englische Sprache übersetzt werden
- Mehrere Datenkorpusse, welche für Textzusammenfassungen verwendet werden können, wurden evaluiert und überarbeitet
- Verschiedene neuronale Netzwerkarchitekturen zur Erstellung von Textzusammenfassungen wurden trainiert und evaluiert
- Cluster von NOTAMs wurden generiert
- Ein Web Interface wurde erstellt, auf welchem NOTAMs gekürzt und als Cluster visualisiert dargestellt werden

In mehreren Durchläufen wurden verschiedene Textzusammenfassmodelle analysiert. Das Ergebnis zeigt auf, dass «Transfer Learning» von englischen Satzverkürzungen auf NOTAMs nicht geeignet ist. Die Verwendung des Word Embeddings «GLOVE», welches globale Vektorrepräsentationen von Wörtern zur Verfügung stellt, verringerte die Performanz des Algorithmus. Daraus erschliesst sich, dass NOTAMs in einer Art verfasst werden, welche starke Differenzen zur englischen Sprache aufweist.

Die erzeugten Cluster besitzen klar merkbare Muster. Themengebiete dieser Cluster können mehrheitlich identifiziert werden. Die überwiegende Zahl der NOTAMs, welche keinem grösseren Themengebiet angehören, befinden sich in Restclusters.

Die erzielten Ergebnisse sind in einem Web Interface visualisiert. In diesem können verschiedene vordefinierte Pakete von NOTAMs ausgewählt werden. Die ausgewählten NOTAMs werden daraufhin zusammengefasst und zu Cluster verarbeitet. Die NOTAMs werden in roher, ausgeschriebener und zusammengefasster Darstellung angezeigt. Die Cluster wurden mithilfe eines interaktiven Scatter Plots visualisiert.

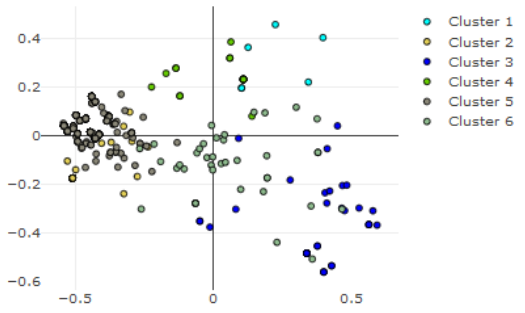

Abbildung 1: Visualisierung von sechs Clustern. Diese beinhalten 400 NOTAMs, welche von Düsseldorf nach München am 5. Mai 2019 aktiv waren. Die Cluster werden durch einen Farbcode unterschieden.

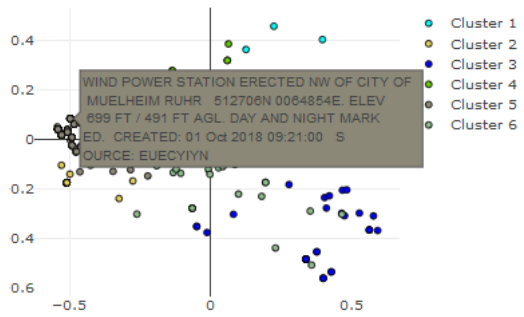

Abbildung 2: Detailansicht eines NOTAMs von Clusters 5. Beschreibt den Standort und die Höhe einer Windkraftanlage.

3. Lösungskonzept

Während der dreiwöchigen Initialisierungsphase des Projektes wurde das benötigte Fachwissen im Bereich «Text Summarization», «Keras» und «NOTAMs» erarbeitet. Darauffolgend wurden Sprints in der Länge von je drei Wochen durchgeführt.

Sowohl für das Training als auch für die Evaluation der Text Zusammenfassung werden NOTAMs benötigt. Nach Anfrage wurde ein Zugriffspunkt zur API der «International Civil Aviation Organization» gewährt. Mithilfe dieser API konnten regelmässig alle aktiven NOTAMs gesammelt werden.

Es existiert kein frei zugänglicher Datensatz, welcher NOTAMs mit einer zugehörigen gekürzten Form anbietet, weshalb die Technik «Transfer learning» verwendet wurde. Bei diesem Verfahren wird ein Datensatz zum Training verwendet, welcher sich nicht in der gleichen Domäne wie der Datensatz zur Evaluation befindet. Als Datensatz mit der höchsten Performanz stellte sich «Compressed Sentences» von Google heraus. Dieser beinhaltet 100’000 ausgeschriebene englische Sätze kombiniert mit den jeweiligen Kurzfassungen.

Bei der Evaluation der möglichen Textverarbeitungsalgorithmen fiel der Entscheid des ersten Versuches auf einen Ansatz, welcher extraktive und abstraktive Textzusammenfassung kombiniert. Da es nicht gelang, das Modell auf die Aufgabenstellung dieser Bachelorarbeit anzupassen, wurde der Versuch abgebrochen.

Daraufhin wurden mehrere neuronale Netzwerkarchitekturen zur abstraktiven Textzusammenfassung verwendet. Keras Modelle wurden auf einem GPU Rechner via Docker Container deployed und trainiert. Die daraus resultierenden Gewichtungen des Netzes wurden persistiert und können wiederverwendet werden. Die trainierten Netzwerke wurden schlussendlich auf einem Evaluationsset von NOTAMs ausgewertet.

Für das Clustering wurde eine «Term frequency/inverse document frequency (TF/IDF)» Matrix basierend auf den gesammelten NOTAMs erstellt. Diese Matrix wurde verwendet, um «unsupervised Clustering» mithilfe von «kMeans» durchzuführen. Die Dimensionen der erhaltenen Cluster wurden mit «Principal component analysis (PCA)» reduziert, um das Ergebnis im 2D-Raum darstellen zu können. Zur Evaluation dieser Cluster wurde deren Kohäsion untersucht und ein Interview mit Experten durchgeführt. Weitere Versuche zum Clustering unter der Verwendung von Word Embeddings waren ursprünglich geplant. Da diese Embeddings jedoch eine tiefe Performanz bei der Kürzung der NOTAMs aufwiesen, wurde von diesem Versuch abgesehen.

Schlussendlich wurde ein Web UI basierend auf «HTML», «css», «JavaScript» und «Flask» implementiert. Dieses wurde auf einer Virtual Machine des Enterprise Labs deployed und ist innerhalb des HSLU Netzwerkes erreichbar.

4. Spezielle Herausforderungen

NOTAMs werden auf verschiedenen Plattformen im Internet angeboten, es existieren jedoch keine Sammlungen von nicht aktiven NOTAMs. Zusätzlich sind die öffentlich zugänglichen Plattformen nicht geeignet, um grosse Mengen an NOTAMs zu sammeln. Aus diesem Grund wurde Kontakt mit der UN Organisation «International Civil Aviation Organization» aufgenommen. Diese gewährte einen zeitbefristeten Zugang zu Ihrer API, auf welcher direkt auf alle aktiven NOTAMs zugegriffen werden kann.

NOTAMs und deren Verfassungsweise werden von verschiedenen Entitäten reguliert, was dazu führt, dass die verwendeten Abkürzungen nicht einheitlich sind. Ausserdem existieren Abkürzungen, welche mehrere Definitionen besitzen. Mehrere Listen von Abkürzungen wurden überarbeitet und zusammengefügt, um eine möglichst akkurate Übersetzung der NOTAMs sicherzustellen. Mehrdeutige Abkürzungen wurden entfernt.

5. Ausblick

Bei dem Trainieren der «Text Summarization» Modellen wurde «Transfer learning» verwendet, da kein Datensatz mit gekürzten NOTAMs vorhanden ist. Falls ein solcher Datensatz zukünftig öffentlich gemacht wird, könnte dieser verwendet werden, um die Textkürzungsmodelle zu trainieren. Ansonsten ist es möglich, manuell einen Datensatz zu erstellen.

Abstraktive Textzusammenfassmodelle schreiben einen neuen Text, welcher den Inhalt des Gesamttext gekürzt repräsentieren soll. Dies kann zu fehlerhaften Änderungen führen. Numerische Werte wie Koordinaten und Zeitangaben werden oft nicht identisch übernommen, was zu einer Verfälschung der ursprünglichen Informationen führt. Dies könnte umgangen werden, indem diese kritischen Informationen durch Platzhalter ersetzt werden. Falls ein solcher Platzhalter im gekürzten NOTAM eingesetzt wird, würde dieser durch die ursprünglichen Informationen ersetzt werden.

Die verwendeten Modelle führen eine abstraktive Textzusammenfassung durch. Ein Versuch mit einem extraktiven Textzusammenfassmodell würde weitere Einblicke liefern. Es ist jedoch unwahrscheinlich, dass «Transfer learning» auf einem extraktiven Modell eine höhere Performanz aufweist als auf dem abstraktiven Modell.