Domänen-sensitives Temporal Tagging für Event-zentriertes Information Retrieval[1]

Jannik Strötgen[2]

Abstract: Da Zeit- und Ortsinformationen in beinahe allen Kontexten eine bedeutende Rolle spielen, kommen sie in Form von Zeit- und Ortsausdrücken häufig in Texten vor. Oft werden dort solche Ausdrücke benutzt, um auf etwas zu referenzieren, das irgendwann irgendwo stattfand, stattfindet, oder stattfinden wird – also um auf Events zu verweisen. Bis jetzt werden *Event-bezogene Informationsbedürfnisse* von Standardansätzen des Information Retrievals jedoch bei weitem nicht hinreichend abgedeckt. Im Rahmen der vorgestellten Dissertation wurden neuartige Frameworks entwickelt, mit denen Dokumentensammlungen in Bezug auf zeitliche, räumliche und Event-bezogene Informationen durchsucht und exploriert werden können. Eine besonders wichtige Rolle spielt dabei auch *HeidelTime*, ein domänen-berücksichtigendes, mehrsprachiges System zum Erkennen und Normalisieren von Zeitausdrücken, das im Rahmen dieser Arbeit entstanden ist und für sämtliche Domänen und Sprachen Evaluationsergebnisse erzielt, die den aktuellen Stand der Forschung widerspiegeln.

1 Einführung

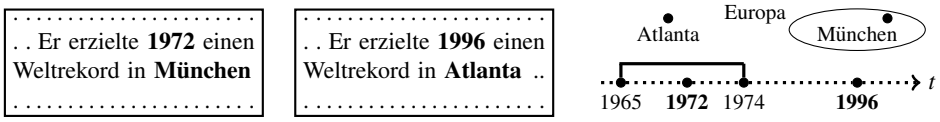
Suchmaschinen wie Google oder Bing werden benutzt, um für ein bestimmtes Informationsbedürfnis, das Nutzer durch eine Suchanfrage ausdrücken, Dokumentensammlungen (z.B. das Internet) zu durchsuchen und Dokumente in nach Relevanz geordneten Ergebnislisten zu erhalten. Dabei werden diverse Informationen genutzt, um die Relevanz der Dokumente zu errechnen, z.B. der textuelle Inhalt, aber auch die Beliebtheit von Webseiten sowie Nutzerfeedback. Die Motivation des Themas der vorgestellten Dissertation liegt ebenfalls im Bereich des Information Retrievals. Dabei spielen zwei Konzepte beim Analysieren des textuellen Inhalts von Dokumenten eine zentrale Rolle: Raum und Zeit.

Räumliche und zeitliche Informationsbedürfnisse sind allgegenwärtig. So wurde in mehreren Studien gezeigt, dass viele Internetsuchanfragen räumliche und zeitliche Terme enthalten [NRD08, Zh06]. Außerdem sind räumliche und zeitliche Ausdrücke in Texten aller Art omnipräsent. Zum Beispiel werden Nachrichtentexte in der Regel an einem bestimmten Tag veröffentlicht und beschreiben, was an diesem Tag oder kurz davor bzw. danach geschehen ist oder wird. Somit kommen zeitliche Ausdrücke wie *heute*, *morgen*, *letzte Woche* oder *11. März* oft vor. Ähnliches gilt für geographische Ausdrücke wie Ortsnamen. Allerdings gilt dies nicht nur für Nachrichtentexte, sondern für Texte aller Art, beispielsweise Biographien und Dokumente über historische Geschehnisse.

Bei näherer Betrachtung erkennt man schnell, dass räumliche und zeitliche Ausdrücke nicht isoliert vorkommen, sondern häufig benutzt werden, um zu beschreiben, was an ei-

[1] Englischer Titel der Dissertation: „Domain-sensitive Temporal Tagging for Event-centric Information Retrieval“ [St15].

[2] Max-Planck-Institut für Informatik, Saarbrücken, jannik.stroetgen@mpi-inf.mpg.de



(a) Dokument A. (b) Dokument B. (c) Hilfreiches Wissen.

Abb. 1: Beispiel für Herausforderungen beim Event-zentrierten Information Retrieval.

nem bestimmten Ort zu einer bestimmten Zeit passiert, also um auf *Events* zu referenzieren. Unabhängig davon, ob man eine Dokumentensammlung nach zeitlichen, räumlichen oder Eventinformationen durchsuchen möchte, ist es elementar, dass der Inhalt von Dokumenten nicht schlicht als Menge von Termen betrachtet wird, sondern deren Semantik verstanden wird. So ist es wichtig, dass eine Suchmaschine nicht nur erkennt, dass in einem Dokument ein Wort wie *morgen* vorkommt, sondern auch „weiß“, was es bedeutet.

Angenommen ein Nutzer interessiert sich für Weltrekorde in einer bestimmten Region zu einer bestimmten Zeit. Für die Suchanfrage *„Weltrekord in Europa zwischen 1965 und 1974“* soll eine Suchmaschine Dokumente bewerten, von denen für zwei jeweils ein kurzer Auszug in Abb. 1(a) und 1(b) dargestellt ist. Betrachtet man nur die vorkommenden Wörter, sind beide Dokumente offensichtlich gleich relevant. Ist allerdings bekannt, dass *München* in Europa und *1972* im Zeitintervall [1965,1974] liegen (vgl. Abb. 1(c)), kann Dokument A klar als relevanter bestimmt werden. Noch schwieriger wird eine Bewertung, wenn im Text statt mit *1972* mit *10 Jahre später* auf das gleiche Jahr verwiesen würde.

Die der Dissertation zugrundeliegende Hypothese lässt sich wie folgt formulieren: (i) Wenn in Dokumenten vorkommende zeitliche und räumliche Ausdrücke erkannt und in ein Standardformat normalisiert werden, (ii) wenn Suchmaschinen einfach zu nutzende Möglichkeiten bieten, zusätzlich zu textuellen auch räumliche und zeitliche Bedingungen zu formulieren, und (iii) wenn Suchmaschinen Wissen über die hierarchische Organisation von Zeit- und Rauminformation zur Verfügung steht, dann lassen sich zeitliche, räumliche und Event-zentrierte Informationsbedürfnisse bedienen. Um diese Hypothese zu prüfen und zu bestätigen, wurden im Rahmen der Dissertation zahlreiche Beiträge geleistet:

- Entwurf und Implementierung von *HeidelTime*, einem Temporal Tagger zum Extrahieren und Normalisieren von Zeitausdrücken aus Texten verschiedener Domänen und Sprachen, dessen Qualität den aktuellen Forschungsstand widerspiegelt,
- Entwicklung des Konzeptes sogenannter *Raum-Zeit-Events* sowie die Analyse zahlreicher Verfahren ihrer Extraktion aus Texten,
- Formalisierung eines mehrdimensionalen Querymodells zur Kombination zeitlicher, räumlicher und textueller Bedingungen sowie Realisierung eines Rankingansatzes für solche Anfragen, der zusätzlich zwei Arten von Näheinformationen einbezieht,
- Frameworks zum Explorieren von aus Texten extrahierten Eventinformationen sowie Entwicklung eines Modells zur Bestimmung von Dokumentenähnlichkeiten, das allein auf zeitlichen und räumlichen Information beruht.

2 Wichtige Charakteristika Zeitlicher & Räumlicher Informationen

Es gibt drei Charakteristika zeitlicher und räumlicher Informationen, aufgrund derer diese Informationen für zahlreiche Such- und Explorationsaufgaben äußerst nützlich sind.

(i) Zeitliche und räumliche Informationen sind wohldefiniert: Gegeben zwei Zeitpunkte oder -intervalle oder zwei Ortspunkte oder -regionen, dann lässt sich ihre Beziehung identifizieren, z.B. als zeitliche Relation wie „vor", „nach" oder „überlappend" [Al83] oder als räumliche Relation wie „innerhalb", „überlappend"oder „unverbunden" [Co97].

(ii) Zeitliche und räumliche Informationen sind normalisierbar: Unabhängig der verwendeten Terme oder Sprache können zwei Ausdrücke, die auf den gleichen Ort bzw. die gleiche Zeit referenzieren, die gleichen normalisierten Werte in einem Standardformat zugewiesen bekommen.

(iii) Zeitliche und räumliche Informationen lassen sich hierarchisch organisieren: Sowohl zeitliche als auch geographische Information können verschiedene Granularitäten aufweisen. Da feingranulare Orte oder Zeiten häufig innerhalb gröbergranularer Orte und Zeiten liegen, existieren Zeit- und Raumhierarchien.

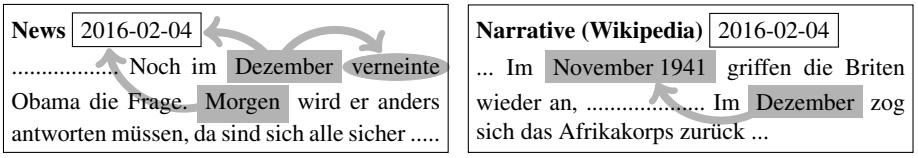
Auch für die wissenschaftlichen Beträge der hier beschriebenen Dissertation spielen diese Charakteristika eine zentrale Rolle, wie in den folgenden Kapiteln deutlich wird.

3 Temporal Tagging mit HeidelTime

Die zwei Hauptaufgaben eines Temporal Taggers sind das Erkennen von Zeitausdrücken in Texten sowie ihre Normalisierung zu Werten in einem Standardformat. Bevor jedoch der im Kontext der Dissertation entwickelte Temporal Tagger HeidelTime vorgestellt wird, werden zunächst wichtige Informationen zu Zeitausdrücken im Allgemeinen aufgezeigt, Schwachstellen verwandter Arbeiten genannt und die Wichtigkeit erläutert, warum Temporal Tagger Texte verschiedener Domänen unterschiedlich handhaben sollten.

Verschiedene Arten und Vorkommnisse von Zeitausdrücken in Texten
Zum Annotieren von Zeitausdrücken in Texten hat sich in den letzten Jahren *TimeML* (temporal markup language) als Standard herauskristallisiert [Pu10] und fast alle Ansätze zum Temporal Tagging versuchen Zeitausdrücke TimeML-folgend zu adressieren. In TimeML werden sie in vier Kategorien eingeteilt: Date, Time, Duration und Set. Zeitausdrücke, die auf einen Zeitpunkt referenzieren, fallen in die Kategorien Date (wenn die Granularität mindestens ein Tag ist) und Time (Ausdrücke feinerer Granularität). Beispiele sind *März 2010* bzw. *9 Uhr morgens*. Ausdrücke der Kategorie Duration beschreiben Zeitintervalle (z.B. *5 Monate*) und solche der Kategorie Set werden verwendet, um wiederkehrende Ereignisse zu beschreiben (z.B. *zweimal pro Woche*).

Des weiteren ist es wichtig, Ausdrücke für Zeitpunkte in Bezug auf ihre textliche Realisierung zu unterscheiden, denn diese hat direkten Einfluss auf die Schwierigkeit einen Ausdruck richtig zu normalisieren. Während in der Literatur bereits viele verschiedene Ka-



(a) Nachrichtenartiger Text (Dezember ist `2015-12`). (b) Narrativer Text (Dezember ist `1941-12`).

Abb. 2: Texte verschiedener Domänen sollten beim Temporal Tagging unterschieden werden. Die Referenzzeit für „Dezember "ist in (a) das Puklikationsdatum und in (b) ein Zeitausdruck in Text.

tegorisierungen vorgeschlagen wurden, werden Zeitausdrücke in der beschriebenen Dissertation in explizite, implizite, relative und unterspezifizierte Ausdrücke aufgeteilt. Diese vier Kategorien spiegeln unmittelbar die Schwierigkeit einer korrekten Normalisierung wider. Zur Normalisierung expliziter Ausdrücke bedarf es keiner Kontextinformationen (z.B. *Mai 1999*, `1999-05`) und für implizite Ausdrücke reicht etwas Wissen aus, das nicht dem standardmäßigen Zeitwissen entspricht (z.B. zum Normalisieren von *Columbus Day 2013*, dass Columbus Day immer der zweite Montag im Oktober ist). Im Gegensatz dazu benötigt man zum Normalisieren von relativen (*gestern*, *ein Jahr später*) und unterspezifizierten Ausdrücken (*Montag*, *März*) eine Referenzzeit und für letztere zusätzlich die Relation zur Referenzzeit. Beides können – je nach Textsorte – schwer zu bestimmende Informationen sein, die jedoch für eine erfolgreiche Disambiguierung nötig sind.

Schwachstellen verwandter Arbeiten & domänenspezifische Herausforderungen
Das Hauptproblem voriger Arbeiten zu Temporal Tagging ist, dass Texte unabhängig ihrer Eigenschaften alle gleich verarbeitet wurden. Für das erfolgreiche Extrahieren und Normalisieren von Zeitausdrücken sollten Texte verschiedener Domänen jedoch unterschieden werden. In Abb. 2 ist ein Hauptproblem dargestellt. Die Referenzzeit für relative und unterspezifizierte Ausdrücke ist in Nachrichtentexten häufig das Publikationsdatum (Abb. 2(a)), aber in narrativen Texten muss sie im Text selbst identifiziert werden (Abb. 2(b)).

In einer breit angelegten Untersuchung von Korpora, die vier verschiedenen Domänen angehören (Nachrichten, Wikipedia, SMS und wissenschaftliche Publikationen zu klinischen Studien), wurden in der hier beschriebenen Dissertation zahlreiche weitere Charakteristika erkannt, die für erfolgreiches Temporal Tagging höchste Bedeutung haben. Während sie in bisherigen Ansätzen zum Temporal Tagging keine Berücksichtigung fanden, wurden nun zahlreiche Strategien für verschiedene Domänen entwickelt und in HeidelTime realisiert.

Ein weiteres Problem bisheriger Arbeiten ist, dass häufig entweder nur Englisch oder wenige weitere Sprachen adressiert wurden und bisherige Systeme nur schwer für andere Sprachen erweiterbar sind.

HeidelTime – Anforderungen & Design
Um die im Rahmen der Dissertation durchgeführte Forschung zu Event-zentriertem Information Retrieval nicht nur auf englischsprachigen Nachrichtentexten durchführen zu können, bedurfte es eines neuen Temporal Taggers. Somit wurde HeidelTime als ein System entwickelt, das folgenden Anforderungen gerecht wird:

(i) Extraktion und Normalisierung sollen von hoher Qualität sein,
(ii) hochqualitative Ergebnisse für Texte unterschiedlicher Domänen sollen erzielt werden,
(iii) zusätzliche Sprachen sollen problemlos einzubinden sein,
(iv) neue Module sollen einfach zu integrieren sein (z.B. für weitere implizite Ausdrücke),
(v) wenn nötig sollen Anpassungen und Erweiterungen einfach möglich sein.

Um die Anforderungen zu adressieren, wurde HeidelTime als regelbasiertes System entwickelt. Die Extraktion basiert auf regulären Ausdrücken, zusätzlichen linguistischen Informationen wie Wortarten, und der Möglichkeit Ressourcen (z.B. Wortlisten für nichtstandardsprachliche Terme) einzubinden. Für die Normalisierung sind domänenspezifische Strategien zur Referenzzeiterkennung realisiert, und es werden linguistische Informationen verwendet, z.B. durch Tempuserkennung. Um weitere Sprachen und Module einfach integrieren zu können, liegt zwischen sprachunabhängigen und sprachabhängigen Komponenten eine strikte Trennung vor. So können beispielsweise Ressourcen für neue Sprachen hinzugefügt werden, ohne dass der Quellcode von HeidelTime verändert werden muss.

Neben der einfacheren Erweiterbarkeit hat auch folgende Annahme dazu geführt, HeidelTime als regelbasiertes System zu entwickeln: aufgrund des relativ geschlossenen Vokabulars, mit dem Zeitausdrücke gebildet werden, sowie der Notwendigkeit Zeitausdrücke nicht nur zu erkennen, sondern auch zu normalisieren, haben regelbasierte Systeme gegenüber Machine Learning-basierten Systemen beim Temporal Tagging Vorteile. Diese Annahme konnte in zahlreichen (offiziellen) Evaluierungen belegt werden (siehe unten).

HeidelTime – adressierte Sprachen und Domänen
Während HeidelTime im Rahmen der vorgestellten Dissertation mit Strategien für vier Domänen sowie Ressourcen für zahlreiche Sprachen entwickelt wurde, hat HeidelTimes Design und wohldefinierte Regelsprache dazu geführt, dass zahlreiche externe Forscher HeidelTime nicht nur verwenden, sondern auch erweitert haben, z.B. um weitere Sprachen abzudecken (u.a. französisch [MT14]). HeidelTime unterstützt mittlerweile vier Domänen und enthält manuell erstellte Ressourcen für 13 Sprachen. Für einige davon ist HeidelTime der einzige Temporal Tagger. Außerdem wurde kürzlich eine Erweiterung präsentiert, durch die erste Ressourcen für über 200 weitere Sprachen automatisch generiert wurden [SG15].[3]

HeidelTime – Evaluierungen und Forschungswettbewerbe
Für sämtliche Sprachen und Domänen werden mit HeidelTime auf existierenden sowie im Rahmen der Dissertation erstellten und veröffentlichten Korpora Extraktions- und Normalisierungsergebnisse erzielt, die den aktuellen Stand der Forschung darstellen. Auch in Forschungswettbewerben hat HeidelTime als bestes System abgeschnitten: TempEval-2 (englisch) [Ve10], TempEval-3 (englisch, spanisch) [Uz13], EVALITA (italienisch) [Ca14]. Zusätzlich wurde in einer Kreuzevaluation gezeigt, in der HeidelTime mit Einstellungen für vier Domänen auf Korpora verschiedener Textsorten getestet wurde, dass die Berücksichtigung unterschiedlicher Domänen zu signifikaten Verbesserungen führt. Beispielsweise verbessert HeidelTime Temporal Tagging von narrativen Texten (Wikipedia) um etwa 20 Prozentpunkte (Standard-Evaluationsmaß value F1) gegenüber vorigen Ansätzen.

[3] HeidelTime ist frei verfügbar (GPL lizensiert), https://github.com/HeidelTime/heideltime/.

4 Raum-Zeit Events

Während es in der Literatur eine Unmenge an Eventdefinitionen und -konzepten gibt, wurde in der hier vorgestellten Dissertation das Konzept sogenannter Raum-Zeit-Events eingeführt, wonach Events lediglich als Kombination von Raum- und Zeitinformation angesehen werden ($e = \langle t, g \rangle$). Um sie aus Texten normalisiert zu extrahieren, bedarf es neben eines Temporal Taggers eines sogenannten Geotaggers, der geographische Ausdrücke erkennt und normalisiert. Im Vergleich zum Temporal Tagging spielen beim Geotagging Charakteristika verschiedener Domänen eine untergeordnete Rolle und – auch deshalb – existierten bereits gute, frei verfügbare Systeme, auf die zurückgegriffen werden konnte.

Obwohl Raum-Zeit-Events eine sehr vereinfachende Definition von Events sind, ist es doch eine sehr sinnvolle und mächtige. Neben der Tatsache, dass Events häufig zu einer bestimmten Zeit an einem bestimmen Ort stattfinden, können durch diese Betrachtung von Events auch sämtliche Eigenschaften von Orts- und Zeitinformationen auf Events angewandt werden (siehe Kap. 2). Außerdem lassen sie sich bereits mit relativ einfachen Mitteln gut erkennen. Eine Analyse zahlreicher heuristischer und linguistischer Methoden zeigte zwar, dass ein einfacher Ansatz, der alle Kookkurrenzen von geographischen und temporalen Ausdrücken in einem Satz als Events extrahiert, verbessert werden kann. Dennoch erzielt aber bereits der Kookkurrenzansatz gute Ergebnisse, die direkt für komplexere Aufgaben verwendet werden können, also z.B. für Event-zentriertes Information Retrieval. Zudem ist die Normalisierung von Events unabhängig von der Extraktionsmethode, da sie durch den Temporal Tagger und den Geotagger sichergestellt wird.

Für geographisches, temporales, und Event-zentriertes Information Retrieval sowie für verschiedene Event-zentrierte Explorationsszenarien wurden Dokument-Profile sowie einige Konzepte zum „Rechnen“ mit Raum-Zeit-Informationen entwickelt.

- Gegeben eine Dokumentensammlung D, dann sind mit jedem Dokument $d \in D$ ein *temporales*, *geographisches*, und *Event-Dokument-Profil* ($tdp(d)$, $gdp(d)$, $edp(d)$) assoziiert, in denen alle aus d extrahierten Zeitausdrücke, Ortsausdrücke bzw. Events in normalisierter Form vorliegen.
- Mithilfe der *temporalen Vorgängerrelation* $\prec_T$ lässt sich die Relation zweier normalisierter Zeitausdrücke t_i und t_j, $t_i \neq t_j$, bestimmen als $t_i \prec_T t_j$ oder $t_j \prec_T t_i$.
- Mit der *geographischen Unverbundenheitsrelation* $\emptyset_G$ lässt sich formulieren, dass für zwei normalisierte geographische Ausdrücke g_i, g_j gleicher Granularität $g_i \emptyset_G g_j$ gilt, wenn $g_i \neq g_j$.
- Die temporalen und geographischen Mappingfunktionen $\alpha_T(t_i') = t_i''$ und $\alpha_G(g_i') = g_i''$ mappen einen normalisierten Zeit- bzw. Ortsausdruck zur nächstgröberen Granularität einer gegebenen Hierarchie.

Basierend auf den Dokument-Profilen sowie den eingeführten Relationen und Mappingfunktionen lassen sich beliebige Zeitausdrücke, Ortsausdrücke und Events miteinander vergleichen. Dies wurde in der hier vorgestellten Dissertation für zahlreiche Information Retrieval und Explorationsszenarien ausgenutzt.

5 Raum-, Zeit- & Event-zentriertes Information Retrieval

Mit textbasierten Standardansätzen des Information Retrievals lassen sich temporale, geographische und Event-bezogene Informationsbedürfnisse wie das in Kap. 1 eingeführte Beispiel kaum zufriedenstellend formulieren. Außerdem werden typischerweise bezüglich des Inhalts der Texte nur Wörter betrachtet, wodurch beispielsweise ignoriert wird, dass „1972“ und „1996“ Zeitausdrücke und „München“ und „Atlanta“ geographische Ausdrücke sind, die Teil von Intervallen bzw. Regionen sein können. Somit lässt sich das in Abb. 1(c) dargestellte Wissen gar nicht erst nutzen.

Deshalb wurden in der hier vorgestellten Dissertation ein mehrdimensionales Querymodell formalisiert, ein prototypisches graphisches Nuzterinterface für temporale und geographische Anfragen realisiert, und Rankingverfahren entwickelt, die sowohl Zeit- als auch Ortsausdrücke als besondere semantische Konzepte berücksichtigen. Zusätzlich wurden zahlreiche Explorationsszenarien entworfen, wie die gemeinsame Karten-basierte Darstellungen aller Events der relevantesten Dokumente zu einer Event-bezogenen Suchanfrage zur interaktiven Exploration relevanter Eventinformationen. Aus Platzgründen sollen im Folgenden jedoch lediglich die Motivation und Haupteigenschaften des Rankingmodells für multidimensionale Suchanfragen beschrieben sowie das Event-zentrierte Dokumentenähnlichkeitsmodell näher betrachtet werden.

Rankingmodell für temporale und geographische Suchanfragen
Wie in Kap. 1 beschrieben enthalten Suchanfragen häufig temporale und geographische Ausdrücke. Diese werden von Standard-Suchansätzen spärlich berücksichtigt. Im Gegensatz dazu wurden in Forschungsarbeiten Rankingmodelle entworfen, die Raum- und/oder Zeitkomponenten zusätzlich zur textuellen Anfrage zulassen ($q = \langle q_{text}, q_{temp}, q_{geo} \rangle$). Die zeitliche und/oder geographische Relevanz von Dokumenten wird dann zusammen mit Standardrankingverfahren für textuelle Suche in einem finalen Ranking zusammengefasst.

Vorige Arbeiten [Be10, Pu07, MMB09] betrachteten allerdings die einzelnen Komponenten der Suchanfragen unabhängig voneinander. Dies ist jedoch nicht intuitiv, da die textuelle Nähe von Wörtern $w \in q_{text}$ und Zeit- und Ortsausdrücken, die in den Intervallen und Regionen q_{temp} und q_{geo} enthalten sind, ignoriert wird. Deshalb wurde hier ein Rankingmodell entwickelt, das diese Unabhängigkeitsannahme auflöst. Neben der Entwicklung neuer Rankingverfahren für die Zeit- und Ortskomponenten, ist eine weitere Haupteigenschaft des Modells, dass belohnt wird, je näher Terme aus q_{text} und relevante Zeit- und Ortsausdrücke in den Texten vorkommen. In einer breiten Evaluierung wurde gezeigt, dass diese Modelleigenschaft zu verbesserten Rankingergebnissen beiträgt.

Event-zentriertes Modell zur Erkennung von Dokumentenähnlichkeiten
In vielen Such- und Explorationsszenarien können Informationen über Ähnlichkeiten von Dokumenten äußerst hilfreich sein. Allerdings sind Ähnlichkeitsbewertungen häufig subjektiv und Dokumente können in Bezug auf unterschiedlichste Eigenschaften als ähnlich betrachtet werden, z.B. aufgrund ihrer Sprache, ihrer Struktur, ihres Themas, enthaltener Wörter oder auch enthaltener semantischer Konzepte. Somit ist auszuschließen, dass ein einziges Ähnlichkeitsmaß als das beste oder gar als das einzig richtige angesehen werden kann. Allerdings basieren viele existierende Modelle unmittelbar auf den in den Doku-

menten enthaltenen Wörtern. Im Gegensatz dazu wurde hier ein neues Ähnlichkeitsmodell entwickelt, das allein auf aus den Texten extrahierten und normalisierten Events basiert. Dieses Event-zentrierte Ähnlichkeitsmodell ist vollständig unabhängig von den in den Dokumenten vorkommenden Wörtern und somit sogar sprachunabhängig. Das im folgenden erläuterte Modell kann deshalb als Komplement zu existierenden, Wörter-basierten Modellen angesehen werden, das neue Arten von Ähnlichkeiten aufdeckt.

Die Schlüsselidee des Modells ist, dass zunächst die Ähnlichkeiten zwischen allen Events zweier Event-Dokument-Profile $edp(d_1)$ und $edp(d_2)$ (also zweier Dokumente) bestimmt werden, diese dann sinnvoll aggregiert werden, und schließlich eine geeignete Normalisierung durchgeführt wird. Basierend auf den Konzepten der temporalen und geographischen Mappingfunktionen aus Kap. 4 – und somit auf den zugrundeliegenden Zeit- und Ortshierarchien – und unter Berücksichtigung weiterer Bedingungen werden zunächst Eventähnlichkeiten bestimmt. Sei α die Summe aller benötigten Mappingschritte, um Gleichheit zweier Events zu erhalten, sei β die maximale Anzahl der involvierten Werte einer Dimension (Raum/Zeit), und sei α_{poss} die Anzahl nach erfolgreichem Mapping noch möglicher Mappingschritte, dann lässt sich die Ähnlichkeit zweier Events bestimmen als

$$sim_e(e_1, e_2) = \frac{1}{(1+\alpha)} \times (\alpha_{poss} + 1).$$

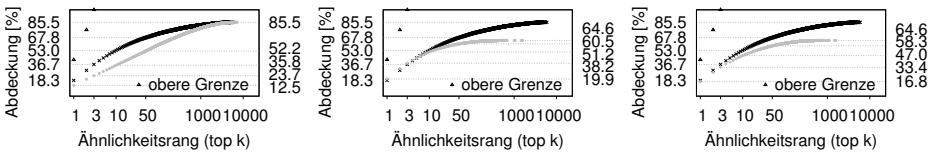
Basierend auf den Ähnlichkeiten aller Events zweier Dokumente und einiger Bedingungen für eine geeignete Kombination dieser Ähnlichkeiten sowie einer sinnvollen Normalisierung, lassen sich Event-zentrierte Ähnlichkeiten zwischen Dokumenten bestimmen mit

$$d\text{-}sim_e(d_1, d_2) = \frac{\sum_{i=0}^{n} \sum_{j=0}^{m} sim_e(e_i, e_j)}{min(n,m)},$$

wobei n, m die Anzahl der Events in $edp(d_1)$, $edp(d_2)$ sind. Aufgrund des Aufbaus des Ähnlichkeitsmodells können drei Komponenten getrennt analysiert werden: *Granularity Mapping (M)*, ohne dass nicht-identische Events stets eine Ähnlichkeit von 0 haben, *Granularity Weighting (W)*, ohne dass die ursprünglichen Granularitäten der Events nicht betrachtet werden, und *Normalization (N)*, ohne dass der aggregierte Ähnlichkeitswert nicht bezüglich der Anzahl der in den Dokumenten vorkommenden Events normalisiert wird.

Für die schwierige Aufgabe einer Evaluierung wurde zusätzlich zu einer manuellen Bewertung ein Verfahren entwickelt, das auf Grundlage eines mehrsprachigen Korpus mit sich entsprechenden Dokumenten (aus Wikipedia extrahierte Dokumente, die über Sprachlinks verknüpft sind) die Güte des Ansatzes im großen Stil bewertet. Als je ähnlicher verlinkte Dokumente erkannt werden, als desto besser wird das System bewertet. In Abb. 3 ist dargestellt, wie häufig in einem viersprachigen Korpus mit knapp 24000 Dokumenten für ein Dokument d_{org}, die sprachverlinkten Dokumente d_{link} im schlechtesten Fall den Rank k einnehmen, also $d\text{-}sim_e(d_{org}, d_{link}) \leq k$. Zusätzlich zum vollständigen Modell (schwarz) werden auch die Ergebnisse des Modells mit nur je einem der drei Eigenschaften dargestellt (jeweils in grau). Dies verdeutlicht die Wichtigkeit aller drei Modellkomponenten.

Offensichtlich können mit dem neuen Modell Ähnlichkeiten zwischen Dokumenten nur berechnet werden, wenn aus beiden Dokumenten Events extrahiert wurden, wodurch sich



(a) Mit Mapping (grau). (b) Mit Weighting (grau). (c) Mit Normalization (grau).

Abb. 3: Evaluierung des Gesamtmodells (schwarz) im Vergleich zu Teilmodellen.

erklären lässt, dass nur 85% der sprachverlinkten Dokumente insgesamt als ähnlich erkannt wurden. Das den in Abb. 3 dargestellten Ergebnissen zugrundeliegende Korpus ist jedoch sehr heterogen und deckt zahlreiche Wikipedia Kategorien ab. Werden stattdessen Dokumente mit typischerweise vielen Events betrachtet (bspw. basierend auf den Wikipediakategorien „Biographie“ und „Geschichte“), sind die erzielten Evaluationsergebnisse noch deutlich höher [St15]. Außerdem konnte mithilfe eines Vergleichs mit einem wortbasierten Ähnlichkeitsmodell gezeigt werden, dass zu einem Großteil mit dem neuen Modell andere Dokumente als ähnlich bestimmt werden, die – wie in einer manuellen Evaluation getestet – als äußerst ähnlich betrachtet werden können.

6 Schlussfolgerungen

In der hier vorgestellten Dissertation wurden die Themen Temporal Tagging sowie temporales, geographisches und Event-zentriertes Information Retrieval adressiert. Durch die Entwicklung von HeidelTime wurde der aktuelle Forschungsstand beim Temporal Tagging vor allem bezüglich Domänenunabhängigkeit und Mehrsprachigkeit auf ein neues Level gehievt. Auch im Bereich des Information Retrievals wurden wertvolle Leistungen erbracht, etwa durch die Formulierung des mehrdimensionalen Querymodells sowie der Entwicklung eines Rankingverfahren, das im Vergleich zu vorigen Ansätzen die unrealistische Unabhängigkeitsannahme zwischen den Dimensionen auflöst. Zusätzlich wurden nützliche Beiträge zur Event-zentrierten Dokumentenexploration geliefert, etwa durch Karten-basierte Ansätze und ein sprachübergreifendes Dokumentenähnlichkeitsmodell.

Literaturverzeichnis

[Al83] Allen, James F.: Maintaining Knowledge about Temporal Intervals. Communications of the ACM, 26(11):832–843, 1983.

[Be10] Berberich, Klaus; Bedathur, Srikanta J.; Alonso, Omar; Weikum, Gerhard: A Language Modeling Approach for Temporal Information Needs. In: ECIR'10. S. 13–25, 2010.

[Ca14] Caselli, Tommaso; Sprugnoli, Rachele; Speranza, Manuela; Monachini, Monica: EVENTI: EValuation of Events and Temporal INformation at Evalita 2014. In: EVALITA'14. 2014.

[Co97] Cohn, Anthony G.; Bennett, Brandon; Gooday, John; Gotts, Nicholas M.: Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. Geoinformatica, 1(3):275–316, 1997.

[MMB09] Machado, Jorge; Martins, Bruno; Borbinha, José: LGTE: Lucene Extensions for Geo-Temporal Information Retrieval. In: GIIW'09. 2009.

[MT14] Moriceau, Véronique; Tannier, Xavier: French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In: LREC'14. S. 3239–3243, 2014.

[NRD08] Nunes, Sérgio; Ribeiro, Cristina; David, Gabriel: Use of Temporal Expressions in Web Search. In: ECIR'08. S. 580–584, 2008.

[Pu07] Purves, Ross S.; Clough, Paul; Jones, Christopher B.; Arampatzis, Avi; Bucher, Benedicte; Finch, David; Fu, Gaihua; Joho, Hideo; Syed, Awase Khirni; Vaid, Subodh; Yang, Bisheng: The Design and Implementation of SPIRIT: a Spatially Aware Search Engine for Information Retrieval on the Internet. International Journal of Geographical Information Science, 21(7):717–745, 2007.

[Pu10] Pustejovsky, James; Lee, Kiyong; Bunt, Harry; Romary, Laurent: ISO-TimeML: An International Standard for Semantic Annotation. In: LREC'10. S. 394–397, 2010.

[SG15] Strötgen, Jannik; Gertz, Michael: A Baseline Temporal Tagger for All Languages. In: EMNLP'15. S. 541–547, 2015.

[St15] Strötgen, Jannik: Domain-sensitive Temporal Tagging for Event-centric Information Retrieval. Dissertation, Institute of Computer Science, Heidelberg University, 2015.

[Uz13] UzZaman, Naushad; Llorens, Hector; Derczynski, Leon; Allen, James F.; Verhagen, Marc; Pustejovsky, James: SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In: SemEval'13. S. 1–9, 2013.

[Ve10] Verhagen, Marc; Saurí, Roser; Caselli, Tommaso; Pustejovsky, James: SemEval-2010 Task 13: TempEval-2. In: SemEval'10. S. 57–62, 2010.

[Zh06] Zhang, Wei Vivian; Rey, Benjamin; Stipp, Eugene; Jones, Rosie: Geomodification in Query Rewriting. In: GIR'06. S. 23–27, 2006.

Jannik Strötgen studierte an der Universität Heidelberg Computerlinguistik und Volkswirtschaftslehre (Magister Artium, 2009). In seiner Magisterarbeit beschäftigte er sich mit heuristischen und linguistischen Methoden zur Extraktion von Relationen zwischen biomedizinischen Entitäten aus wissenschaftlichen Texten. Während seines Studiums arbeitete er als Tutor am Institut für Computerlinguistik, als studentische Hilfskraft beim Fraunhofer Institut für Algorithmen und Wissenschaftliches Rechnen (SCAI) sowie als Intern in der IT Forschungsabteilung von GlaxoSmith-Kline in Pennsylvania. Von 2010 bis 2015 war er Doktorand am Institut für Informatik der Universität Heidelberg, wo er seit 2009 als wissenschaftlicher Mitarbeiter am Lehrstuhl für Datenbanksysteme tätig war. Im März 2015 verteidigte er seine Doktorarbeit per Rigorosum mit summa cum laude. Nach weiteren Monaten an der Universität Heidelberg wechselte er im Oktober 2015 als Postdoc an das Max-Planck-Institut für Informatik (Saarbrücken), wo er sich vor allem den Themen Informationsextraktion und Information Retrieval widmet. Das im Rahmen seiner Dissertation entwickelte System *HeidelTime* ist frei verfügbar und erfreut sich in der Forschungsgemeinschaft großer Beliebtheit.