Tekstuele informatie-extractie: een overzicht

Deze tekst geeft op een toegankelijke manier een overzicht van een aantal concepten, technieken en toepassingen binnen het domein van tekstuele automatische informatie-extractie.

De onderliggende tekst is niet geschreven voor onderzoekers, maar bedoeld op niveau van toepassingen, meer specifiek voor mensen uit de media-sector. De state-of-the-art is daarom niet diepgaand, en de uitleg is kort en eenvoudig. Wel zijn in beperkte mate verwijzingen naar de wetenschappelijke literatuur opgenomen, als startpunt voor verdere verdieping in de beschreven onderwerpen.

Dit document is een van de resultaten van de Unified Thesaurus haalbaarheidsstudie die door VIAA werd georganiseerd[1].

De auteurs van dit document zijn:

Hans Paulussen	iMinds – ITEC – KU Leuven
Sam Coppens	iMinds – MMLab – UGent
Erik Mannens	iMinds – MMLab – UGent
Philip Leroux	iMinds – IBCN – UGent
Thomas Demeester	iMinds – IBCN – UGent

Inhoudstafel

[1] Meer informatie over de Unified Thesaurus haalbaarheidsstudie staat op de iminds.be website: http://www.iminds.be/nl/projecten/2014/09/08/unified-thesaurus.

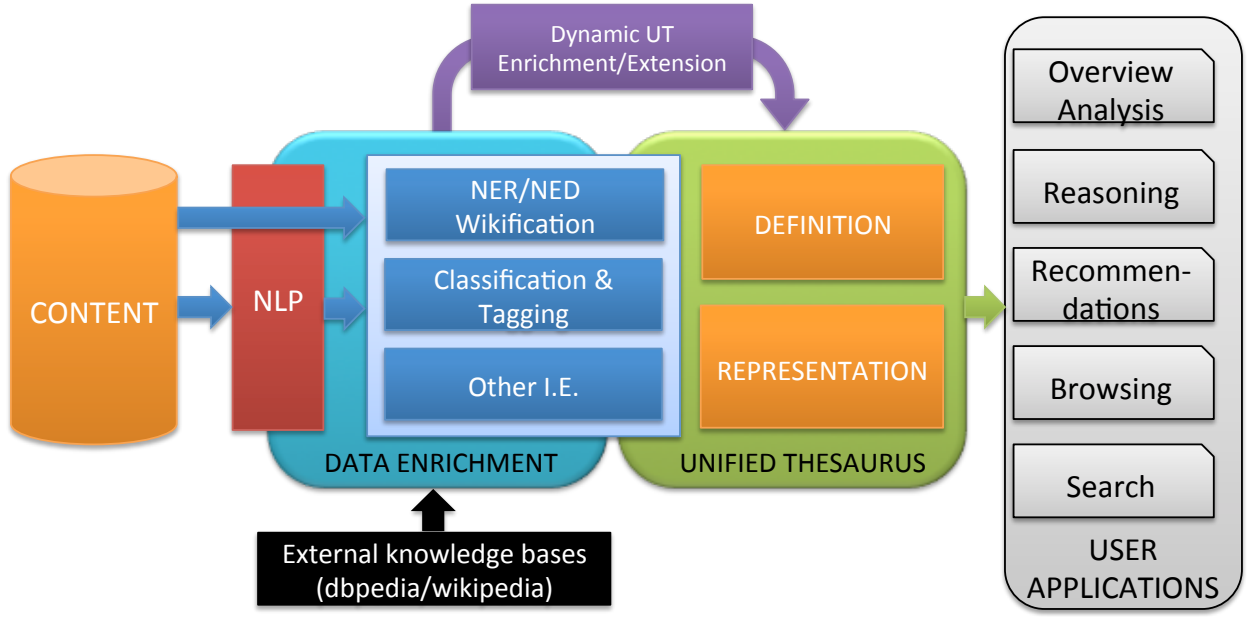
1. Introductie

Binnen de domeinen van data en tekst analyse, natuurlijke taalverwerking (Natural Language Processing of NLP), informatie extractie en verrijking, enz... vallen heel wat definities, termen, technieken en toepassingsdomeinen. De bedoeling van deze taak is om zowel het algemene kader van al deze technieken te schetsen, en tevens meer diepgaande technische informatie te geven m.b.t. de state-of-the-art van een aantal van deze technieken, welke best passen binnen het kader van dit project.

We proberen voornamelijk een goed leesbare tekst te produceren, die informatief is voor de partners in de media-, archief-, en erfgoedsector, als voor de onderzoeksgroepen. Het document wordt hiertoe opgesteld in het Nederlands, en de beschrijvingen blijven eerder kwalitatief dan heel strikt wetenschappelijk. Wel worden telkens referenties voorzien naar meer diepgaande wetenschappelijke literatuur.

2. Situering van informatie-extractie binnen het referentiekader van de studie

In de literatuur bestaan heel wat onderzoeksdomeinen die tot doel hebben om tekst als input te nemen om er daarna “iets” mee te gaan doen, bijvoorbeeld deze te gaan verrijken of in een gestructureerde vorm gieten, wat dan de grondslag moet vormen voor nieuwe toepassingen. Heel vaak zal voor bepaalde toepassingen ook een aaneenschakeling gebeuren van verschillende technieken. Men kan daarbij spreken over een pipeline waarbij alle componenten samen een bepaald eindresultaat opleveren. Voor elke deelcomponent van de pipeline bestaan vaak verschillende technieken om deze te implementeren en ook het aantal componenten aanwezig in de pipeline kan sterk verschillen afhankelijk van de eindtoepassing. Voor de keuze van deze componenten voor een bepaalde toepassing wordt niet alleen de accuraatheid van het eindresultaat in rekening gebracht, maar ook factoren als de computationele kost, onderhoudbaarheid, flexibiliteit, het dynamisch karakter, etc.



Figuur 1: Algemeen overzicht referentie pipeline

2.1. NLP vs Informatie Extractie.

De algemene referentiepipeline die als rode draad binnen het UT initiatief werd gebruikt, is afgebeeld in Figuur 1. In dit document focussen we op de DATA ENRICHMENT component vermits de informatie-extractie taken en systemen zich binnen deze component situeren. Zoals aangegeven op de figuur maken we binnen de referentie architectuur nog onderscheid tussen 2 grote blokken:

- NLP component: Natural Language Processing, deze component zal alle taken omvatten die van een naakte, niet geannoteerde tekst starten om daarop een (taalkundige) basisanalyse uit te voeren. Dit process levert als resultaat een verrijkte versie van de tekst aan. Deze verrijking omvat typisch de toevoeging van taalkundige informatie zoals type woordenaanduiding, zinsstructuur ontleding, lemmatizering, aanduiden van entiteiten, etc.
- I.E. component (Information Extraction): deze component gebruikt typisch de verrijkte data van de NLP stap met als doel om een bepaalde toepassing (vb. categoriseren, annoteren, etc.) uit te voeren.

Wetenschappelijk gezien is een strikt onderscheid tussen NLP en IE eerder artificieel. Basis NLP taken zoals het extraheren van part-of-speech tags, kunnen worden gezien als een vorm van informatie extractie. Hier maken we voor de duidelijkheid wel een onderscheid, en we zien informatie extractie typisch als ruimer en dichter bij de toepassing dan NLP. Tevens worden

taken op document-niveau, zoals classificatie van documenten, soms niet als een echte IE taak beschouwd. Het belangrijkste voor dit document is echter dat het duidelijk is waarover het precies gaat, en daarom gaan we niet verder in op de subtiele verschillen in definities (en in Sectie 3 worden deze naast elkaar behandeld).
Afhankelijk van de toepassing die men beoogt zal men steeds een verschillende combinatie van NLP modules met I.E. modules combineren. Voor iedere toepassing bestaan er per module ook vaak meerdere technieken die een gelijkaardig resultaat beogen, maar met andere voor- en nadelen. Per techniek of tool zullen we steeds aanduiden welke functionaliteit deze omvatten (via afkortingen).

Voor NLP componenten bestaat de functionaliteit typisch uit 1 of meerdere van volgende taken:

- Part of Speech (POS) tagging, d.w.z., annoteren van woordsoorten
- lemmatisatie: woorden hervormen naar hun basisvorm
- Named Entitiy Recognition: het extraheren van entiteiten (i.e. personen, locaties of organisaties uit een tekstdocument

Bij de verschillende I.E. componenten bestaat de functionaliteit typisch uit 1 of meerdere van volgende taken:

- het uniek identificeren van eigennamen, via een link naar unieke identiteiten in een knowledge base: entity linking of Named Entity Desambiguation (NED), bijvoorbeeld naar Wikipedia identiteiten, als een onderdeel van wikificatie. Merk op dat wikificatie bijvoorbeeld ook kan worden gedaan op contextuele keywords, in plaats van enkel eigennamen.
- het groeperen van teksten
    - via classificatie: binnen een van buitenaf opgelegde categorisatie.
    - via clustering: onderverdeling via logische samenhang, gebaseerd op bepaalde kenmerken van de data zelf.
- het bepalen van contextuele keywords.
- het bepalen van sentiment.
- het detecteren van gebeurtenissen of evenementen.

2.2. Implementatie van componenten

Machinaal leren vs. handgeschreven regels of heuristieken

Moderne NLP en I.E. algoritmen zijn meestal gebaseerd op machinaal leren (machine learning, ML). Het paradigma van ML verschilt fundamenteel van de meeste voorgaande pogingen om NLP te doen, waarbij vaak grote hoeveelheden hand-gecodeerde regels of heuristieken werden gebruikt, wat tot logge en moeilijk onderhoudbare systemen kon leiden. In het vakgebied van statistische ML worden algemene leeralgoritmen gebruikt, om automatisch (in een training-fase)

via de analyse van grote corpora van echte data soortgelijke (hoewel vaak intuitief niet interpreteerbare) regels te gaan leren.

Supervised learning vs unsupervised learning

Een belangrijk onderscheid dat kan worden gemaakt binnen de gebruikte Machine Learning technieken, is het onderscheid tussen Supervised en Unsupervised technieken. Supervised learning is het leren van een bepaalde functionaliteit uit vooraf (meestal handmatig) geannoteerde trainingsvoorbeelden. Een supervised learning algoritme analyseert de training data, en leert een bepaalde functie, die kan worden gebruikt om predicties te doen voor niet geziene situaties. Typische classificatie taken worden bijvoorbeeld vaak via supervised learning geïmplementeerd.

Non-supervised learning gaat niet uit van training data, en analyseert typisch zeer grote hoeveelheden data. Door het ontbreken van training data, is het resultaat van unsupervised algoritmen typisch niet rechtstreeks bruikbaar voor toepassingen, maar door het vermogen om zeer grote hoeveelheden data te verwerken (veel meer dan ooit manueel zou kunnen worden geannoteerd), kunnen supervised algoritmen met de resultaten wel krachtiger worden gemaakt. Clustering is typisch een unsupervised taak (hoewel er ook supervised varianten bestaan), met als voorbeeld het gebruikt van woord clusters (gebaseerd op een groot corpus), om Named Entity Recognition meer accuraat te maken.

3. NLP & I.E. technieken

Onderstaande paragrafen beschrijven de hedendaagse technieken die worden gebruikt voor een aantal NLP en I.E. taken. Zoals reeds aangehaald, bestaan er voor bepaalde toepassingen meerdere benaderingen die naast elkaar kunnen worden gebruikt. Voor elk van de onderstaande onderdelen, plaatsen we de internationale state-of-the-art voorop, met toelichting van de status voor het Nederlands, enkele toepassingsgebieden, en tenslotte een beperkt aantal referenties naar meer gespecialiseerde literatuur. Tevens plaatsen we verschillende benaderingen naast elkaar, met bijhorende voor- en nadelen.

3.1. POS tagging

Part-of-speech (PoS) tagging is gewoonlijk de eerste vorm van tekstannotatie. Nadat een tekst is opgesplitst (of gesegmenteerd) in zinnen en in tokens (woorden en niet-woordvormen (bijv. leestekens)), worden alle tokens geannoteerd met een PoS tag, een label dat de woordsoort

aanduidt. De PoS tag vormt de sleutel tot verdere annotatie, in het bijzonder voor parsing en named entity recognition.

Omschrijving technieken

Zoals voor elke fase in taaltechnologie wordt algemeen een onderscheid gemaakt tussen regelgebaseerde en statistische taggers. De eerste vereist heel wat werk voor het uitschrijven van regels die de context beschrijven waarin een bepaalde woordsoort voor een bepaald woord wordt geselecteerd. Bovendien is die methode niet robuust genoeg voor woordsoortcombinaties die buiten de voorziene regels vallen. De statistische aanpak gaat uit van probabiliteitsberekeningen: hoe waarschijnlijk heeft een bepaald woord in een bepaalde context een bepaalde woordsoort. Een statistische tagger wordt hiervoor getraind op basis van grote tekstsamples of tekstcorpora. Er wordt een onderscheid gemaakt tussen supervised en non-supervised training. In het eerste geval wordt eerst een testcorpus handmatig geannoteerd, zodat de woord-woordsoort-combinaties de nodige trainingsinformatie opleveren. In het tweede geval worden niet geannoteerde tekstsamples gebruikt. De kwaliteit van stochastische taggers hangt sterk af van de hoeveelheid tekstdata die voor de training van de tagger wordt gebruikt. De codes van de woordsoorten bestaan gewoonlijk uit twee delen: de hoofdcategorie (bijv. werkwoord) en de subcategorie (bijv. geslacht en getal). De verdere specificering van de subcategorieën hangt af van de morfologische structuur van de taal of van de toepassing die men voor ogen heeft.

State-of-the-art

Op dit ogenblik gaat de voorkeur meestal uit naar statistische taggers, vooral omdat de ontwikkeling, in verhouding tot regelgebaseerde taggers, minder tijd vraagt. Niettegenstaande dat zijn er ook een aantal taggers die features van beide of andere type taggers bevatten. Enkele bekende taggers zijn: Brill (Brill 1992), TNT (Brants 2000), TreeTagger (Schmid 1994).

Status voor Nederlandse taal

Voor het Nederlands zijn verschillende taggers ontwikkeld, waaronder: MBT tagger, CGN tagger fabriek, MXPost, PAROLE tagger (Daelemans & Strik 2002). Als algemene referentie wordt tegenwoordig vooral Frog aangehaald als representatieve tagger specifiek voor het Nederlands. Frog is de integratie van een aantal memory based NLP modules, ontwikkeld aan de universiteit van Tilburg (Van den Bosch et al. 2007). De meest gebruikte tagset voor het Nederlands is die ontwikkeld werd voor het CGN (Corpus Gesproken Nederlands) (Van Eynde et al. 2000).

Typische toepassingsgebieden

De annotatie van PoS tags vormt de basis voor vele toepassingen in de taaltechnologie. Een typisch voorbeeld is de ontwikkeling van spellingcheckers of het samenstellen van een basiswoordenschat op basis van de meest frequente woorden uit een representatief corpus. De

woordsoort en het lemma vormt ook de sleutel voor allerlei toepassingen waar verdere informatie over woorden in een lexicon moeten worden opgezocht.

Referenties

- Brants, T. (2000), TnT - A Statistical Part-of-Speech Tagger, "Proc 6th Applied Natural Language Processing Conference", ANLP-200
- Brill, Eric (1992), A simple rule-based part of speech tagger. In Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155.
- Daelemans, W. & H. Strik (red.) (2002), “Het Nederlands in de taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen”. Een rapport in opdracht van de Nederlandse Taalunie.
- Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. "Proceedings of International Conference on New Methods in Language Processing", Manchester, UK.
- Van den Bosch, A., Busser, G.J., Daelemans, W., and Canisius, S. (2007). An efficiënt memory-based morphosyntactic tagger and parser for Dutch, In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), "Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting", Leuven, Belgium, pp. 99-114.
- Van Eynde, F., J. Zavrel, and W. Daelemans (2000). Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus. In "Proceedings of the second International Conference on Language Resources and Evaluation" (LREC-2000), pages 1427–1434, Athens, Greece, 2000.

3.2. Parsing

Met parsen bedoelt men automatische zinsontleding. Een zin bestaat uit woorden die woordgroepen vormen en een functie in de zin hebben. Een parse tree of een ontledingsboom is een boomstructuur die de verbanden tussen de verschillende zinsdelen of constituenten toont. Door middel van parsing probeert men dus de zinsdelen en de relatie tussen die zinsdelen op te sporen. Algemeen wordt een onderscheid gemaakt tussen constituent en dependency parsing. Daarnaast wordt ook een onderscheid gemaakt tussen full parsing en shallow parsing.

Omschrijving technieken

Voor de ontwikkeling van een parser maakt men algemeen gebruik van een grammatica die de syntactische structuren van de taal beschrijft en een lexicon dat de mogelijke morfosyntactische klassen van een woord opgeeft. De parser tracht dan de zinnen te analyseren overeenkomstig de regels van de grammatica. De ontwikkeling van de grammatica is een arbeidsintensieve opdracht. Naast deze deductieve methode wordt ook meer en meer gebruik gemaakt van zogenaamde treebanks, die een verzameling syntactisch geannoteerde teksten bevatten. Dankzij de syntactische annotatie kan een grammatica op inductieve manier uit een treebank worden afgeleid, waardoor de grammaticaregels niet handmatig moeten worden opgesteld.
Voor het parsen maken we een onderscheid tussen full parsing en chunk parsing (of shallow parsing). In het eerste geval probeert men een volledige syntactische boom op te bouwen, in het tweede geval beperkt men zich tot het selecteren van typisch clusters. Bij full parsing wordt gebruik gemaakt van grammatica formalismen gebaseerd op boomstructuren (o.a. LFG, CCG, HPSG en TAG). De verschillende theorieën gaan terug naar de ontwikkeling in de generatieve grammatica (Chomsky 1956). Afhankelijk van het bewandelde pad in de boomstructuur spreekt men van twee soorten parsers: bottom-up of top-down. Een veel gebruikt formalisme in statistische parsers is PCFG (probabilistic context free grammar) wat een probabilistisch model is van een contextvrije grammatica.
Bij full parsing denkt men vooral aan constituency parsing. Daarnaast maakt men ook gebruik van o.a. dependency parsing. Het laatste is meer en meer in trek, omdat het flexibeler is voor het herkennen van woordgroepen, en daardoor gemakkelijker resultaten oplevert. Constituentenknopen (bijv. NP, VP, AP) die typisch zijn voor constituenten grammatica's, komen niet voor in dependency grammars. Door de flexibiliteit is dependency grammar ook interessant voor talen waar woordvolgorde minder vast is. Een aantal parsers maken gebruik van zowel c-parsing als d-parsing, waarbij het niet duidelijk is of de ene afgeleid is van de. Malt-parser is een "zuivere" d-parser, maar Stanford geeft output van zowel d-trees als c-trees.
Via chunk parsing (Abney 1994) probeert men belangrijk woorgroepen op te sporen. Het betreft dan vooral nominale, verbale en prepositionele constituenten. Het zijn vooral de nominale constituenten die belangrijk zijn bij het opsporen van named entities. Indien een volledige parse niet noodzakelijk is, dan is chunk parsing een betere oplossing om snel de gewenste knopen op te sporen.

State-of-the-art

Vooral voor het Engels zijn een aantal parsers ontwikkeld op basis van een handmatig opgestelde grammatica, waaronder SRI core language engine en XTAG grammatica op basis van Tree Adjoining Grammar (TAG). Een bekend voorbeeld van depency parsing is MaltParser (Nivre 2003). Belangrijke constituent parsers zijn de Stanford (Klein en Manning 2003) en de Berkeley (Petrov et al. 2006) parsers. Beide parsers maken gebruik van het PCFG formalisme. Er zijn een aantal chunk parsers beschikbaar, meestal specifiek voor het Engels, en dikwijls geïntegreerd in NLP toolkits: o.a. Illinois chunk parser, Apache OpenNLP, GATE.

Status voor Nederlandse taal

Voor het Nederlands werden een aantal parsers ontwikkeld: o.a. Amazon-casus, Carper Technologies, Corrie (Daelemans en Strik 2002). Sinds de STEVIN projecten (Spyns en Odijk 2013) is duidelijk ALPINO (van Noord 2006) het meest bekend. Voor chunk parsing zijn ook andere tools beschikbaar, waaronder Frog.

Typische toepassingsgebieden

Parsers worden gebruikt voor bijv. correctie van grammaticale fouten, waarbij de verbanden tussen woorden noodzakelijk is (bijv. congruentie tussen onderwerp en werkwoord). Parsers worden ook ingezet voor de analyse van vragen in information retrieval. Shallow parsing is ook belangrijk voor het opsporen van named entities.

Referenties

- Abney, Steven (1991), "Parsing By Chunks", Principle-Based Parsing, Kluwer Academic Publishers, pp. 257–278.
- Klein, Dan and Christopher D. Manning (2003), Accurate Unlexicalized Parsing. in Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.
- Nivre, J. (2003). An Efficiënt Algorithm for Projective Dependency Parsing. In Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03), Nancy, France, 23-25 April 2003, pp. 149-160.
- Petrov, Slav, Leon Barrett, Romain Thibaux and Dan Klein (2006), " Learning Accurate, Compact, and Interpretable Tree Annotation" in COLING-ACL 2006
- Spyns, P. En J. Odijk (eds.) (2013), Essential speech and language technology for Dutch, results by the STEVIN programme. Springer.
- van Noord, Gertjan (2006). "At Last Parsing Is Now Operational" In: TALN 2006, pp. 20-42.

3.3. NER

Named entity recognition (NER) of naamherkenning is een techniek om eigennamen te herkennen in teksten. Het betreft zowel losse woorden als woordgroepen. De herkenning gebeurt door analyse van de woorden en de context van de woorden. Er zijn zowel regelgebaseerde als statistische manieren voor naamherkenning beschikbaar.

Omschrijving technieken

NER (Named entity recognition) of naamherkenning is een techniek om named entities op te sporen. Named entities zijn o.a. eigennamen die verwijzen naar personen, organisaties of plaatsnamen. Eenvoudig gezegd gaat het om encyclopedische woorden die je niet in een gewoon woorden-boek verwacht. Er zijn verschillende manieren om namen te herkennen. In principe geraak je al ver met het herkennen van patronen, zoals woorden die beginnen met hoofdletters, en het gebruik van namenlijsten (ook gazetteers genoemd), maar vrij snel wordt duidelijk dat die manier van werken vrij beperkte resultaten oplevert. Taaltechnologie kan ook hierin een belangrijke rol spelen, al is het maar om potentiële namen op te sporen. Dankzij een goede annotatie van woordsoorten en het (shallow) parsen van teksten kan men gemakkelijker zelfstandige naamwoorden en woordgroepen opsporen, die typisch in aanmerking komen voor named entities. Net zoals bij andere toe-pas-singen van taaltechnologie, wordt ook hier een onderscheid gemaakt tussen regel-gebaseerde en statistische modellen. Bij het laatste wordt gebruik gemaakt van supervised en non-supervised methodes. Om efficiënter te zoeken, wordt niet alleen gebruik gemaakt van de karakteristieken van het te onderzoeken woord zelf. Er wordt ook rekening gehouden met de context van de woorden en het domein waartoe de tekst behoort.

Naast het herkennen van een named entity, is het ook belangrijk de NE te categoriseren, wat een basisstap is om de betekenis van de NE te desambigueren (wat gebeurt binnen NED: named entity disambiguation). Binnen de enamex classificatie maakt men een onderscheid tussen PER (persoon), LOC (locatie), en ORG (organisatie). Het type MISC (miscellaneous) wordt beschouwd als een restcategorie.

State-of-the-art

In statische modellen wordt o.a. gebruik gemaakt van volgende methodes: memory-based learning (MBL), conditional random fields (CRF) en support vector machines (SVM).

Apache Open NLP gebruikt zowel regelgebaseerde als statische procedures voor NER. De Stanford NLP-tools (Finkel et al. 2005) maken gebruik van een CRF NER module. NLP platforms, zoals GATE en OpenCalais, bieden NER aan voor meerdere talen.

NER kwam in de belangstelling dankzij het 6e Message Understanding Conference (MUC-6) (Zie ook Marsh en Perzanowski 1998). Zoals gewoonlijk werd het eerst voor Engels gebruikt. Nederlands werd later ook geanalyseerd (Tjong Kim Sang en De Meulder 2003). Voor een overzicht van gebruikte technieken, verwijzen we naar volgende artikels: Nadeau & Sekine (2007), Mansouri, et al (2008), Samet and Labatut (2013)

NER lijkt gemakkelijk, maar er blijven toch heel wat uitdagingen over, die vooral te maken hebben met het domeinspecifieke karakter van de teksten. Ook de dubbelzinnigheid van woorden speelt hierin een rol: bijv. sommige namen komen overeen met gewone woorden (bijv. Karel Appel), sommige namen kunnen onder verschillende types worden geklasseerd, afhankelijk van de context (bijv. Europa is plaats en organisatie).

Status voor Nederlandse taal

De meeste NER tools zijn taalonafhankelijk, maar het vraagt voor elke taal en elk domein een specifieke training van de tools. De bekendste tools hebben ook een taalmodel voor het Nederlands, maar meestal is het vrij beperkt getrained.
Tools die specifiek voor het Nederlands zijn gemaakt, zijn de volgende.
Binnen het iReadplus project maakt NER deel uit van de NLP pipeline (Paulussen et al. 2014). Daarnaast heeft Zeticon een NER tool ontwikkeld (Van den Bossche et al. 2012), gebaseerd op de CRF techniek, en verrijkt door gebruikt te maken van unsupervised features getrokken uit grote hoeveelheden Nederlandstalige data. Tenslotte omvat ook het product van Newz.nl de NER functionaliteit.

Typische toepassingsgebieden

NER wordt vooral gezien als de basis om extra informatie over personen, organisaties en locaties te linken aan de NE. Het is de sleutel tot verdere semantische analyse van teksten, wat bijvoorbeeld de mogelijkheid biedt om teksten gemakkelijker automatisch te classificeren. Tevens vormt het de sleutel tot de verrijking van teksten, waarbij extra informatie (bijv. uit Wikipedia) aan de named entity kan worden gelinkt.
NER speelt ook een belangrijke rol in faceted search, waar de automatisch gedetecteerde eigennamen toelaten om flexibel door de collecties te browsen.

Referenties

- Atdağ, Samet & Vincent Labatut (2013), A Comparison of Named Entity Recognition Tools Applied to Biographical Texts, "2nd International Conference on Systems and Computer Science", Villeneuve d'Ascq (FR), 228-233.
- Finkel, Jenny Rose, Trond Grenager, & Christopher Manning (2005), Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. "Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics" (ACL 2005), pp. 363-370.
- Mansouri, Alireza, Lilly Suriani Affendey & Ali Mamat (2008), Named Entity Recognition Approaches, in “IJCSNS International Journal of Computer Science and Network Security”, VOL.8 No.2, February 2008, 339-344.
- Marsh, Elaine & Dennis Perzanowski (1998), "MUC-7 Evaluation of IE Technology: Overview of Results", 29 April 1998.

- Nadeau, D. & Sekine, S. (2007), A survey of named entity recognition and classification. “Linguisticae Investigationes”, 30, 3-26.
- Paulussen, Hans, Pedro Debevere, Francisco Bonachela Capdevila, Maribel Montero Perez, Martin Vanbrabant, Wesley De Neve & Stefan De Wannemacker (2014), Building an NLP pipeline within a digital publishing workflow, presented at Computational Linguistics in The Netherlands (CLIN), Leiden, 17 January, 2014.
- Tjong Kim Sang, Erik F. & De Meulder Fien (2003), “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition,” in CONLL ’03 Proceedings of the 7th Conference on Natural Language Learning, vol. 4, Stroudsburg, PA, 2003, pp. 142-147.
- Van Den Bossche, Bruno, Brecht Vermeulen, Johannes Deleu, Thomas Demeester & Piet Demeester (2012), “MediaHaven : Multimedia Asset Management with Integrated NER and Categorization.” In 12th Dutch-Belgian Information Retrieval Workshop, Abstracts, 85–86.

3.4. NED

Omschrijving Technieken

Named Entitye Disambiguation (NED) is de procedure om named entities te linken aan kennisbanken. Het volstaat bijv. niet om Michael Jackson te herkennen als persoon. Je moet ook het onderscheid kunnen maken tussen de muzikant en de bierexpert. Via NED probeert men de dubbelzinnigheid van de named entity te disambigueren, zodat de juiste informatie in bijv. Wikipedia kan worden opgezocht en getoond. Oorspronkelijk werd Named Entity Disambiguation (*NED*) bekomen door entiteitvermeldingen (“entity *mentions”*) die naar dezelfde entiteit refereren, te groeperen.

Door de ontwikkeling van grote kennisbanken (*knowledge base*, KB) wordt Named Entity Disambiguation tegenwoordig bekomen door een entity mention te linken met het overeenkomstig record in een kennisbank. Deze techniek wordt meestal aangeduid met de term *entity linking* (andere mogelijke termen zijn *entity resolution*, *entity reconciliation* en *record linkage*). In de literatuur wordt als kennisbank vaak gebruikgemaakt van Wikipedia omdat deze een enorme hoeveelheid beschrijvingen bevat van zogenaamde “real world named entities”. In dit geval wordt in plaats van de term *entity linking* vaak *Wikificatie* (*Wikification*) of *D2W* (*Disambiguation to Wikipedia*) gebruikt. De oorspronkelijke techniek waarbij enkel een groepering werd uitgevoerd valt onder de term *(cross-document) coreference resolution.*

NED is een uitdagende taak en dit door verschillende aspecten. Een entiteit kan verschillende gangbare vermeldingsvormen (*surface forms*) hebben (Bvb., een organisatie kan aangeduid worden met een afkorting of een voluit geschreven naam). Verschillende entiteiten kunnen ook aangeduid worden met dezelfde surface form (Bvb., personen met identiek dezelfde naam). In de literatuur worden deze aspecten vaak omschreven als *name variation problem* en *name ambiguity problem,* respectievelijk. Een laatste aspect is dat het mogelijk is dat een entiteit dat vermeld wordt in een document geen overeenkomstig record heeft in de KB. In dat geval wordt typisch een nieuwe 'identiteit' gecreëerd, waarmee dan alle verder voorkomende surface forms van die entiteit worden geassocieerd.
De meeste NED technieken maken gebruik van de output van een NER systeem dat reeds de entiteiten detecteert en classificeert.

State-of-the-art

De meeste algoritmes bekomen disambiguatie door het combineren van a priori kennis (bvb., gegeven een surface form, welke entiteit zal volgens de kennisbank hiermee hoogst waarschijnljk bedoeld worden) en contextuele informatie. Verschillende modellen en methodes zijn in gebruik zoals bvb. bag-of-words (BOW), graafgebaseerde methodes, methodes gebruikmakend van machinaal leren (meestal gesuperviseerd, zoals bvb. *Support Vector Machines (SVM)*), enz.

De SOTA op het gebied van NED valt het meest eenvoudig te bepalen door te kijken naar de deelnemende systemen aan de Entity-Linking track van de Text Analysis Conference [6]. Nagenoeg alle NED systemen hanteren hetzelfde algemene schema van document annotatie gevolgd door kandidaat selectie, gevold door het scoren van deze kandidaten. Er bestaat wel veel variatie in hoe deze verschillende stappen concreet geïmplementeerd worden, alsook welke informatie gebruikt wordt voor het desambigueren. Zo gebruikt het best scorende systeem voor de 2012 editie [3] als enige geo-coördinaten om aldus bij de desambiguatie van locaties rekening te kunnen houden met de onderlinge afstanden tussen alle mogelijke kandidaten. Verder laten zij de beslissing over welke vorm nu net te desambigueren afhangen van mogelijke desambiguaties van andere mentions in het document (bv. de keuze tussen "Alexandria", of "library of Alexandria"). [4] houdt de vorm van de mentions dan weer vast, en lost een Markov Random Field op over alle mentions en hun bijhorende kandidaten, daar waar [5, 6] eerst mentions clusteren, vervolgens alls mentions onafhankelijk van elkaar desambigueren, en tot slot een majority vote over de clusters hanteren om tot een desambiguatie op cluster niveau te komen.

Status voor Nederlandse taal

Voor zover wij weten bestaat er momenteel geen commerciële software specifiek voor NED op de Nederlandse taal. In de literatuur ligt de focus vooral op het disambigueren van

Engelstalige teksten. Echter, de ontwikkelde algoritmes kunnen in principe ook toegepast worden op Nederlandstalige teksten. Dit werd o.a. geillustreerd in het DBpedia Spotlight project dat meerdere talen ondersteunt.

Ook binnen het iRead+ project werd een pipeline ontwikkeld waar NED kan uitgevoerd worden op zowel Nederlandstalige als Franstalige teksten. Binnen iRead+ wordt NED bekomen door een surface form te linken met een Semantic Web resource.

Binnen het BEAMER project werd een variant voor het Nederlands ontwikkeld van het model [Mertens et al., 2013] dat werd ontworpen in het kader van de internationale TAC challenge, maar dan gebaseerd op het NER systeem van Zeticon. Hierbij ligt de focus op het verbinden van surface forms met Wikipedia, en het onderling clusteren van surface forms die niet in de knowlege base voorkomen.

Typische toepassingsgebieden

NED kan in heel wat toepassingsgebeiden van nut zijn. Zo kan het een rijkere leeservaring bieden aan lezers en taalleerders op een digitaal platform. Het kan de performantie van (interne) zoeksystemen verbeteren. Het kan een belangrijk onderdeel zijn van tagging/categorisatie-systemen, enz.

Referenties

- [1] L. Mertens, T. Demeester, J. Deleu, P. Demeester and C. Develder. 2012. UGent Participation in the TAC 2012 Entity-Linking Task, Proceedings of the Fifth Text Analysis Conference
- [2] L. Mertens, T. Demeester, J. Deleu and C. Develder. 2013 [TBP]. UGent Participation in the TAC 2013 Entity-Linking Task, Proceedings of the Sixth Text Analysis Conference
- [3] S. Cucerzan. 2012. The MSR System for Entity Linking at TAC 2012, Proceedings of the Fifth Text Analysis Conference
- [4] P. McNamee, V. Stoyanov, J. Mayfield, et al. 2012. HLTCOE Participation at TAC2012: Entity Linking and Cold Start Knowledge Base Construction, Proceedings of the Fifth Text Analysis Conference
- [5] S. Monahan, J. Lehmann, T. Nyberg, J. Plymalle and A. Jung. 2011. Cross-Lingual Cross-Document Coreference with Entity Linking, Proceedings of the Fourth Text Analysis Conference
- [6] S. Monahan and D. Carpenter. 2012. Lorify: A Knowledge Base from Scratch, Proceedings of the Fifth Text Analysis Conference
- [6] http://www.nist.gov/tac/

3.5. Document Classificatie

Omschrijving technieken

Het doel van document classificatie is om documenten toe te wijzen aan 1 of meerdere klassen of categorieën, v.b. ‘economie’, ‘politiek’, etc. In het geval van een mapping op categorieën spreekt men ook van categorisatie. Categorisatie is typisch een supervised taak waarbij men het systeem zal trainen aan de hand van een reeks voorbeeld documenten die reeds manueel aan een correcte categorie zijn toegekend. Nieuwe documenten kunnen daarna automatisch worden toegekend aan deze categorieën.

In de literatuur zijn reeds heel wat document classificatie systemen beschreven, typisch gebaseerd op een heel wat verschillende technieken die te situeren zijn onder de machine learning toepassingen. Zo maakt een van de meest succesvolle methodes bijvoorbeeld gebruik van Support Vector Machines [1]. De 2 grote nadelen van deze technieken zijn de grote hoeveelheid aan trainingsdata die nodig is en het beperkte aantal categorieën waarvoor een voldoende hoeveelheid aan trainingsdata een afdoende nauwkeurigheid kan opleveren.

Een alternatief voor supervised leren is het gebruik van een bestaande ontologie of categorieboom, zoals de Wikipedia categorieboom, met als doel zeer fijne categorisatie uit te voeren [2]. Het probleem met dit type van aanpak is dat de categorieën niet direct gelinkt zijn met de inhoud van de beschouwde collectie. Voldoende accuraatheid verkrijgen vereist daarbij typisch heel wat tweaken en het toevoegen van heuristieken. Daarnaast is ook een externe ontologie vaak niet even snel up-to-date als het eigen nieuwsarchief, hetgeen ook kan leiden tot mismatches tussen beide.

Naast de supervised classificatie taken, worden topic models gebruikt om thematische informatie te ontdekken in grote corpussen van documenten of om het abstracte “onderwerp” te ontdekken in collecties van documenten. Topic models zijn gebaseerd op het idee dat documenten in een corpus kunnen aanzien worden als een mix van onderwerpen, waarbij een onderwerp op zich een probabiliteitsdistributie van de aanwezige woorden is. Het doel is dus om elk document in een corpus te omvatten in een korte beschrijving, om zodoende efficiënt maar nog steeds zo correct mogelijk grote collecties te kunnen verwerken met het oog op basistaken zoals classificatie, zoeken van gelijkaardige artikels, het genereren van samenvattingen, etc.

Topic models zijn typisch een unsupervised techniek wat dus inhoudt dat er geen labeling of annotatie op voorhand dient te gebeuren. Een nadeel hiervan is dat topic models vaak wel zeer moeilijk intuitief te begrijpen zijn.

State-of-the-art

De momenteel algemeen aanvaarde standaard techniek voor topic modellering is door gebruik te maken van Latent Dirichlet Allocation (LDA), dit generatief model werd in 2003 door Blei [3] geintroduceerd. Sindsdien zijn nog verschillende alternatieve topic models ontworpen [4,5,7], vaak als een variatie op LDA. Er bestaan ook supervised topic models [6], waarbij het model wordt geholpen bij de keuze van de best passende topics, via annotaties van een deel van de documenten.

Status voor Nederlandse taal

Binnen het BEAMER project werd voor een eerste prototype ontwikkeld rond automatische categorisatie, in samenwerking met Zeticon, en op data van Mediargus. Verder hebben we niet onmiddellijk kennis van software die specifiek voor het Nederlands een volledig geautomatiseerde categorisatie aanbiedt.

Typische toepassingsgebieden

Het classificeren van media items is van uitzonderlijk belang naar de consument toe, en eveneens voor een vlot intern beheer van de data. Voor de verwerking van grote corpora niet-gestructureerde teksten, worden automatische classificatie technieken gebruikt, die typisch getraind worden via annotaties van een kleine sample van documenten, en vaak krachtiger worden gemaakt via unsupervised features, zoals automatisch gedetecteerde topics.

Referenties

- [1] Zhang, W., Yoshida, T., Tang, X (2008) Text classification based on multi-word with support vector machine, Knowledge-Based Systems, Volume 21, Issue 8, Dec. 2008, 879-886.
- [2] Janik, M. and Kochut. K. J. (2008). Wikipedia in Action: Ontological Knowledge in Text Categorization. In Proceedings of the 2008 IEEE International Conference on Semantic Computing (ICSC '08).
- [3] Blei, D.; Ng, A.; Jordan, M. (2003) 'Latent dirichlet allocation', J. Mach. Learn. Res., vol. 3, pp. 993–1022.
- [4] Blei, D.; Lafferty, J. (2006). 'Correlated topic models', Advances in Neural Information Processing Systems, 18.
- [5] Li, W.; McCallum, A. (2006), 'Pachinko allocation:dag-structured mixture models of topic correlations', International Conference on Machine Learning(ICML).
- [6] Blei, D.; McAuliffe, J. (2007) 'Supervised topic models', Advances in Neural Information Processing Systems.

- [7] Diaz-Aviles, E., Georgescu, M., Stewart, A., and Nejdl, W. (2010). LDA for on-the-fly auto tagging. In Proceedings of the fourth ACM conference on Recommender systems (RecSys '10). ACM, New York, NY, USA.

3.6. Detectie van gebeurtenissen

Omschrijving technieken

De bestaande technieken die gebruikt worden om automatisch gebeurtenissen (events) te detecteren in grote corpora of news streams, kunnen opgesplitst worden in data-gebaseerde en kennisgebaseerde technieken [1].
De eerste soort werkt op basis van statistische relaties tussen verschillende woorden in een corpus, zoals hun frequenties en co-frequenties in verschillende artikels. Hierbij zijn er dus veel artikels nodig om de invloed van ruis te beperken en zinvolle resultaten te krijgen.
De tweede soort (kennisgebaseerde) gebeurtenisdetectie heeft minder data nodig en gebruikt kennis (bv. kennis uit semantische lexicons als WordNet etc.). Op basis van deze kennis worden dan lexico-semantische of lexico-syntactische regels gegenereerd om events te detecteren. Een regel kan hierbij heel simpelweg als volgt uitzien:“aanslag gepleegd in [STAD]”. Het voordeel is dus dat events sneller kunnen worden gedetecteerd en dat hiervoor veel minder data noodzakelijk is. Verder kan bij deze manier van event-detectie ook automatisch data geëxtraheerd worden (bv. soort event, locatie van event etc.). Het nadeel is dat er bij deze technieken veelal veel manueel werk nodig is en dat deze technieken minder goed schalen.
Er bestaan eveneens technieken die een combinatie zijn van beide soorten.

State-of-the-art

De data-gebaseerde technieken werden voor het eerst voorgesteld in het werk rond Topic Detection and Tracking (TDT) [2]. De huidige state-of-the-art bouwt dan meestal ook verder op deze technieken. Zo zijn er technieken die de event-detectie verbeteren door het combineren van verschillende data. In [3] worden verschillende corpora (artikelsets) gecombineerd, terwijl in [4] verschillende voorstellingen van dezelfde artikels worden gecombineerd. Anderzijds wordt er ook geëxperimenteerd met het aanpassen van de voorstelling van de artikels. Zo wordt er gebruikgemaakt van de ‘burstiness’ van de termen (m.a.w., hoe sterk ze een piek in de tijd vormen) om een verbeterde voorstelling te verkrijgen in [5,6].
Bij de kennis-gebaseerde technieken wordt er veel gebruik gemaakt van ontologieën [7] of worden gazetteers [8] gebruikt voor het genereren van de regels. Verder wordt er ook veel onderzoek gedaan naar technieken die unsupervised deze verschillende regels/patronen leren [9].

Status voor Nederlandse taal

In het Nederlands is er veel minder kennis (bv. semantische lexicons) voorhanden dan voor het Engels. Hierdoor wordt er meer gebruik gemaakt van de data-gebaseerde technieken. Wij hebben geen weet van bedrijven die echt inzetten op event detectie. Echter, Oxynade.com is een lokaal bedrijf dat de mogelijkheid biedt aan consumenten om events in te brengen en te omschrijven, waarop ze na automatische classificatie van het event-type een service aanbieden, gebaseerd op deze data.

Typische toepassingsgebieden

Event-detectie wordt gebruikt voor het automatisch detecteren van events. Dit is nuttig voor mediagroepen die op deze manier hun artikels automatisch kunnen bundelen, grafisch voorstellen in functie van de tijd, en om het doorzoeken van hun artikelset te vereenvoudigen.

Referenties

- [1] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. An overview of event extraction
- from text. In Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE
- 2011) at Tenth International Semantic Web Conference (ISWC 2011), volume 779, pages 48–57, 2011.
- [2] James Allan. Topic detection and tracking: event-based information organization, volume 12. Springer, 2002.
- [3] Wang, X., Zhai, C., Hu, X. & Sproat, R. 2007. Mining correlated bursty topic patterns from coordinated text streams. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’07. 784, ACM Press, New York, New York, USA.
- [4] De Smet, W. & Moens, M.-F. 2009. An aspect based document representation for event clustering. In Proceedings of the 19th Meeting of Computational Linguistics in the Netherlands.
- [5] Zhao, W., Chen, R., Fan, K., Yan, H. & Li, X. 2012. A novel burst-based text representation model for scalable event detection. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, (July), 43–47.
- [6] He, Q., Chang, K., Lim, E. & Zhang, J. 2007. Bursty feature representation for clustering text streams. Proc. SIAM Conference on Data Mining, pp. 491–496.
- [7] Frasincar, F., Borsje, J., Levering, L.: A SemanticWeb-Based Approach for Build- ing Personalized News Services. International Journal of E-Business Research 5(3), 35–53 (2009)

- [8] Li, F., Sheng, H., Zhang, D.: Event Pattern Discovery from the Stock Market Bulletin. In: 5th International Conference on Discovery Science (DS 2002). Lecture Notes in Computer Science, vol. 2534, pp. 35–49. Springer-Verlag Berlin Heidelberg (2002)
- [9] Xu, F., Uszkoreit, H., Li, H.: Automatic Event and Relation Detection with Seeds of Varying Complexity. In: AAAI Workshop on Event Extraction and Synthesis (2006)

3.7. Sentiment predictie

Omschrijving technieken

Traditioneel is de bedoeling van sentiment predictie (opinie extractie, sentiment analyse), het bepalen van de algemene contextuele polariteit van een document (artikel, review, blog...) of het (subjectief) standpunt van de auteur ten aanzien van een bepaalde topic. Recent wordt ook onderzoek gedaan naar het detecteren van sentiment dat vanwege het onderwerp door de lezer wordt gehecht aan eerder objectieve data zoals nieuws artikels.
Veel van de hiertoe voorgestelde technieken maken sterk gebruik van lexica met sentiment-houdende termen en uitdrukkingen, bv. in de vorm van een lange lijst woorden met hun a priori polariteit (positief / negatief / neutraal). Een aantal van de belangrijkste problemen is de invloed van de context, waardoor a priori eerder negatieve woorden in de tekst zelf als neutraal of zelfs positief worden aangevoeld.

State-of-the-art

Reed vanaf de jaren 2000 wordt er geprobeerd om automatisch subjectiviteit in tektsten te identificeren [Wiebe et al., 2005], sterk gericht op eigenschappen van de gebruikte termen. Ook werd er onderzoek gedaan om in een affectieve context de polariteit (positief/negatief) te gaan voorspellen, e.g., [Turney, 2002]. Wat we momenteel met sentiment detectie bedoelen, is de beide samen: het identificeren van subjectieve tekst en de bijhorende sentiment polariteit. Een mooi overzicht van het onderzoek hierrond tot 2010 wordt gegeven in [Liu, 2010].

Momenteel wordt dergelijk onderzoek verder uitgewerkt op verschillende soorten tekstueel materiaal, zoals reviews van klanten of gebruikers (over films, producten...) [Thet et al., 2009], [Zhai et al., 2011], sociale media [Tan et al., 2011], blogs [Melville et al., 2009], of zelfs beurs-gerelateerd nieuws [Mizumoto et al., 2012].

Daarnaast is ook de detectie van sentiment uit objectieve nieuws archieven en belangrijk onderzoeksonderwerp aan het worden, i.h.b. voor media-monitoring. Een voorbeeld is een

politieke partij die wenst te weten met welke polariteit zij het vaakst wordt vermeld. De belangrijkste uitdagingen zijn hierbij het ontbreken van sterk sentiment-geladen termen en de subjectiviteit van het sentiment t.a.v. de lezer. Bestaand werk in die richting, bv. [Wilson et al., 2009] legt de focus op de analyse van aparte zinnen, of beperkt de focus tot zeer specifieke domeinen, zoals bv. artikels over bedrijfsovernames [Devitt and Ahmad, 2007]. Ander werk voorspelt de polariteit, afhankelijk van het specifiek standpunt in de tekst, gebaseerd op ontologieën [Scholz and Conrad, 2012]. Algemeen worden meestal basistechnieken uit de NLP gebruikt om geschikte features te genereren, en Machine Learning technieken voor de predicties.

Status voor Nederlandse taal

In het Nederlands zijn bestaande tools die sentiment detecteren heel beperkt, en dat zijn dan vooral platformen die monitoring en analyse van sociale media voorzien. Zo zijn er coosto.com, en obi4wan.nl. Deze laatste integreert naast sociale media ook nieuwsbronnen, fora, en weblogs, en voorziet een totaal sentiment rond de klant in functie van de tijd.

Typische toepassingsgebieden

De toepassingsmogelijkheden volgen rechtstreeks uit de taak zelf. Vaak gaat het om monitoring en analyse van grote hoeveelheden data, om een idee te krijgen van sentiment aangaande specifieke entiteiten (zoals films, bedrijven…).

Referenties

- Wiebe J., and Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005) (Invited paper.)
- Turney, P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 417–424.
- Liu, B. (2010). Sentiment analysis and subjectivity. In Handbook of Natural Language Processing, Second Edition, N. Indurkhya and F. J. Damerau, Eds. CRC Press.
- Thet, T., Na, J., Khoo, C., and Shakthikumar, S. (2009). Sentiment analysis of movie reviews on discussion boards using a linguistic approach. In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (TSA '09). ACM, New York, NY, USA, 81-84.

- Zhai, Z., Liu, B., Xu, H., and Jia, P. (2011). Clustering product features for opinion mining. In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11). ACM, New York, NY, USA, 347-354.
- Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., and Li, P. (2011). User-level sentiment analysis incorporating social networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11). ACM, New York, NY, USA, 1397-1405.
- Melville, P., Gryc, W., and Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09). ACM, New York, NY, USA, 1275-1284.
- Mizumoto, K., Yanagimoto, H., and Yoshioka, M. (2012). Sentiment Analysis of Stock Market News with Semi-supervised Learning. In Proceedings of the 2012 IEEE/ACIS 11th International Conference on Computer and Information Science (ICIS '12). IEEE Computer Society, Washington, DC, USA, 325-328.
- Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics, 35(3):399–433.
- Devitt, A. and Ahmad, K. (2007). Sentiment polarity identification in financial news: A cohesion-based approach. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.
- Scholz, T., and Conrad, S. (2012). Integrating Viewpoints into Newspaper Opinion Mining for a Media Response Analysis. Proceedings of KONVENS 2012, Vienna.

3.8. Detectie van topic-gerelateerde keywords

Omschrijving technieken

Bestaande methodes voor extractie van keywords kunnen opgesplitst worden in gesuperviseerde (supervised) en ongesuperviseerde (unsupervised) technieken [1]. Gesuperviseerde methodes maken gebruik van documenten met reeds toegewezen keywords om eigenschappen van de keywords te herkennen, en zo uit nieuwe documenten keywords te leren extraheren [2]. Anderszijds gebruiken ongesuperviseerde methoden geen reeds geannoteerde documenten, maar concentreren zich op woord-frequentie en het samen voorkomen van woorden [3]. De meest populaire techniek is TF-IDF, waarbij woorden een hogere score krijgen naarmate ze frequent voorkomen in het document maar niet in de volledige collectie [4]. In ander werk worden documenten voorgesteld als (on)gerichte grafen en worden woorden geclusterd op basis van verwantschap, of wordt het PageRank algoritme op de

graaf uitgevoerd [5]. Op basis van de uitkomst van deze algoritmen worden woorden gerangschikt naar relevantie met het document zelf, en de best scorende worden geselecteerd als keywords.

State-of-the-art

In de state-of-the art voor gesuperviseerde methoden wordt steeds op zoek gegaan naar nieuwe eigenschappen van de woorden of machine learning technieken die het aanleren van keyword detectie verbeteren. Dit kan via semantische informatie uit een lexicon zoals WordNet, of met een externe kennisbron zoals Wikipedia [6,7]. In recent onderzoek naar ongesuperviseerde methoden worden de woord-frequentie-gebaseerde methoden ook gecombineerd met data uit WordNet [8], worden gelijkaardige documenten gebruikt als extra invoer voor de algoritmen [9], of wordt informatie over de thematische samenstelling van het document gebruikt [1].

Status voor Nederlandse taal

Voor de Engelse taal zijn een reeks commerciële producten onder de vorm van Web API beschikbaar, voorbeelden zijn de Yahoo Term Extraction Service, Open Calais, of Zemanta. Voor de Nederlandse taal zijn deze voorlopig nog onbestaande, hoewel bestaande methodes relatief eenvoudig kunnen worden toegepast op Nederlandse documenten. Afhankelijk van de techniek zijn een POS-tagger, geannoteerde data, een nederlandstalige Wikipedia, of WordNet nodig.

Typische toepassingsgebieden

Keywords geven een beknopte samenvatting van een document. Gebruikers kunnen zich baseren op keywords om sneller queries te formuleren en informatie uit een collectie te extraheren. Keywords worden ook gebruikt door bedrijven die content produceren of beheren, om inhoud te organiseren, automatisch te classificeren, of om hun zoekmachines te verbeteren.

Referenties

- [1] Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010). Automatic keyphrase extraction via topic decomposition. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, (October), 366–376. Retrieved from http://dl.acm.org/citation.cfm?id=1870694
- [2] Turney, P. (1999). Learning to extract keyphrases from text. Retrieved from http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=8913245
- [3] Grineva, M., Grinev, M., & Lizorkin, D. (2009). Extracting key terms from noisy and multitheme documents. WWW 2009 MADRID! Track: Semantic/Data Web / Session: Mining for Semantics, 661–670. Retrieved from http://dl.acm.org/citation.cfm?id=1526798

- [4] Frank, E., Paynter, G., & Witten, I. (1999). Domain-specific keyphrase extraction. Retrieved from http://researchcommons.waikato.ac.nz/handle/10289/1508
- [5] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. Proceedings of EMNLP, 85. Retrieved from http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf
- [6] Mihalcea, R., & Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. CIKM'07, November 6–8, 2007, Lisboa, Portugal. Retrieved from http://dl.acm.org/citation.cfm?id=1321475
- [7] Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. Information Processing & Management, 43(6), 1705–1714. doi:10.1016/j.ipm.2007.01.015
- [8] Wang, J., Liu, J., & Wang, C. (2007). Keyword extraction based on pagerank. PAKDD 2007, 857–864. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-71701-0_95
- [9] Wan, X., & Xiao, J. (2008). CollabRank: towards a collaborative approach to single-document keyphrase extraction. Proceedings of the 22nd International Conference on …, (August), 969–976. Retrieved from http://dl.acm.org/citation.cfm?id=1599203