Influential Article Review - Development of an Advanced Digital Extraction Method

Ervin Haynes

Russell Davis

Clayton Huff

This paper examines technology. We present insights from a highly influential paper. Here are the highlights from this paper: The main goal of this study is to build high-precision extractors for entities such as Person and Organization as a good initial seed that can be used for training and learning in machine-learning systems, for the same categories, other categories, and across domains, languages, and applications. The improvement of entities extraction precision also increases the relationships extraction precision, which is particularly important in certain domains (such as intelligence systems, social networking, genetic studies, healthcare, etc.). These increases in precision improve the end users' experience quality in using the extraction system because it lowers the time that users spend for training the system and correcting outputs, focusing more on analyzing the information extracted to make better data-driven decisions. For our overseas readers, we then present the insights from this paper in Spanish, French, Portuguese, and German.

Keywords: Entity extraction, Machine learning, Precision of extraction, Text analytics, Natural language processing

SUMMARY

- The main goal of our study was to build very high-precision entity extractors for the Person and Organization categories that would minimize the noisy output . We used the two specific categories, Person and Organization, they are among the top most heavily utilized categories in information retrieval systems across domains, thus they can be used to further improve the NER system's precision and expand its scope through machine learning.
- First, our method improved the precision of entity extraction of two categories and created a good initial seed that can be used for machine learning in the future. Our proposed method ensures that each extractor is specialized in one and only one category: if the rules of a specific classifier do not recognize an entity, it will be ignored and not extracted at all instead of being misclassified. Even though missing a considerable number of potential entities will lower the extractor recall, we have designed a second step for ML that will counteract this weakness and improve recall .
- Furthermore, these categories are general enough so that the extractors can be used and continuously improve entity extractions on the two aforementioned categories across domains and

applications. The improvement in entity extraction quality will also increase the precision of entity relationship extraction. Noisy data becomes less of a roadblock. The end users of the extractor systems will be spending less time in evaluating the quality of the extracted entities and relationships and spend more time instead in analyzing the information retrieved from the system and make better data-driven decisions. Relationship extraction accuracy is particularly important in certain domains, such as intelligence systems, social networking, genetic studies, healthcare, and the like.

HIGHLY INFLUENTIAL ARTICLE

We used the following article as a basis of our evaluation:

Zaghloul, W., & Trimi, S. (2017). Developing an innovative entity extraction method for unstructured data. International Journal of Quality Innovation, 3(1), 1–10.

This is the link to the publisher's website:
https://jqualityinnovation.springeropen.com/articles/10.1186/s40887-017-0012-y

INTRODUCTION

In today's networked global world, people, goods, data, and knowledge move more widely, quickly, and freely in the speed of the light across the various boundaries. For businesses, governments, and scientific communities, this new environment has brought tremendous opportunities, as well as challenges [1]. Not only are countries' economies more connected and interdependent but also the people, governments, politics, and knowledge [2]. The advances in information and communication technology (ICT) have brought a tremendous increase in the amount of data created and shared (big data), techniques, technologies, and systems to extract value from the data. Data analytics are used for a variety of purposes (business, security and safety, scientific discovery, etc.), domains (biology, medicine, education, etc.), and stakeholders (businesses, governments, scientists, and consumers) [3]. Therefore, extracting information and value from data has become critical for academia, the industry, and governments.

Many institutions and organizations are increasingly gathering intelligence by processing and analyzing massive amounts of data that is in textual format and collected from multiple sources and languages. Processing and analyzing such data, which very often are imperfect, incomplete, and unstructured, have become increasingly difficult. Thus, development of new intelligence and analytics (I&A) technologies and improvement of the quality of data processing and analytics have become the focus of governments, mathematicians, computer scientists, and data analysts. Methodologies, techniques, and practices of extracting value from data, i.e., I&A, have changed with the changes in types of data collected and analyzed. I&A 1.0 analyzed structured content; I&A 2.0 analyzes unstructured (text-based) content; and I&A 3.0, which is in the early stages, focuses on mobile and sensor-based content analysis [4].

The emerging areas for text analytics are (1) information extraction (IE)—automatically extracting structured information from documents; (2) topic model (TM)—discovering the main themes in a large and unstructured collection of documents by using algorithms; (3) opinion mining—access, extract, classify, and understand the opinions expressed in many sources including social networks; sentiment analysis is also used for opinion mining; and (4) question answering (Q&A)—answering factual questions (e.g., IBM's Watson, Apple's Siri, Amazon's Alexa, etc.) based on techniques from statistical natural language processing (NLP), information retrieval (IR), and human-computer interaction (HCI) [4].

The purpose of this study is to add to the current literature on the performance of named entity recognition (NER), the building block for IE systems. Specifically, we build an entity extractor for categories Person and Organization, which are entities that have a finite set of identifiers. The proposed extracting method improves (1) the extraction quality of the system by increasing the precision of entity extraction as a result of the initial extracted entities from our method; our highly precised entities will be

used as a seed for training other machine learning (ML) systems, across domains and languages, thus, not only eliminating the need for manual training but also will expand the seed (through learning) and therefore will continue to increase the precision and use of entity extraction across domains and languages; and (2) the experts' experience and use: experts will spend less time for training the system (as is done by the seed entities created by us, and expended by ML) and also less time on fixing faulty entities and relationships extracted by the systems, thus, instead, will spend their time on using the system's outcome to make better data-driven decision.

This paper is organized as follows: the "Literature review" section presents a review of relevant literature for the theoretical concepts of the study; the "The proposed method" section explains the proposed method for entity extraction; the "Discussion" section provides and discusses the study results; and the "Conclusions" section concludes the study by summarizing the study's contributions, limitations, and future research needs.

CONCLUSION

Contributions

The greatest benefit of our proposed method is that it creates high-precision entity extractors for the Person and Organization categories. These extracted entities will make a highly reliable training set for machine learning algorithms to learn the extraction rules for the two categories, and thus improve the extraction quality (F measure) of all two extractors. The number of documents needed for training is usually very large and would require hundreds of hours from each analyst involved. Specialists' time is not only very expensive but also often difficult to find. Thus, eliminating specialists' time to train systems not only results in huge cost savings and but also increases specialists' efficiency as they spend their time performing their job by using and interpreting the outcome of NER systems, instead of training, checking outcomes, and correcting errors. Improvement in precision in extraction of entities improves the precision of entity relationship extraction, thus minimizing the system users' time spent on searching through graphs and fixing faulty relationships.

Another important benefit of our system is its generalizability: (1) There is no need for knowledge in any specific domain, as the two categories we applied our system to are both general in nature and widely used in many knowledge domains. Identifying a person or organization is similar across domains. Thus, there is no need for domain experts to build their own extraction system. After an initial investment in building a set of good extractors, the extractors can become stable over time through applications and be useful in multiple domains. The method can be further refined through the feedback from users. (2) The proposed method for extracting entities' type of Person and Organization can be effectively applied to any other category of entities that have a finite set of sub-types or identifiers. If the target category is not directly applicable, it is possible to generate similar initial high-precision seed entity lists (as long as the number of seed items is large enough), which can be used to train machine-learning systems to learn the extraction rules. The easiest way to obtain such seed entity lists is through the use of specialized dictionaries specifically built with a subset of the known entity set. For example, a category like Locations could benefit from the use of a specialized dictionary or a gazetteer. A seed list of locations could be extracted from the document set; then the feature surrounding each seed could be used by a ML algorithm to generate extraction rules for finding entities. (3) In our proposed method, there is no need for manually tagged training set of documents in any sort or ML in the first stage, increasing the generalizability of applications (not just domain) of the method across many areas, without the need for major modifications.

Finally, our proposed method could also be easily used in different languages that have similar features to English, especially Germanic languages (such as, German, Dutch, Danish, Norwegian, etc.) and Romance languages (such as, Spanish, French and Italian). Replacing the English list of identifiers for each category with the equivalent list from the target language would yield similar results with little to no complications.

Limitations

The main aim of our study was to improve the precision of entity extraction. The quality of our extraction system, however, depends on the quality of the determiner lists. If the categories do not have a finite number of "members," our method would not achieve similar high-precision results. Creation of such lists requires research and time and could vary from one language to another. This method could be a challenge for very large data intensive systems. It would not be a very difficult task, however, to take an English list of determiners and find the equivalent list in other languages.

Another weakness of our method is that it is not applicable to every possible entity extraction category. A category like Locations could highly benefit from the use of a specialized dictionary or a gazetteer (reference). As discussed earlier, a seed list of locations could be extracted from the document set, and then, the feature surrounding each seed could be used by ML algorithms to generate extraction rules or to find similar entities. Such features could be based on part of speech (POS) tags, grammatical constructs, semantics, and many other possible features. The efforts to create such lists will increase the ability to utilize them across application domains, with little to no changes.

Last limitation in using our proposed method comes from the fact that, because it is aimed at maximizing precision, it could lower the recall as an indirect result. Measuring recall in a system that processes a very large number of documents is always a challenge, and therefore, recall in such systems is usually estimated. Such deficiency could be overcome by accurately measuring the recall based on data sets such as CoNLL03 [22] or other similar datasets. This recall evaluation (not just estimation) on a common data set, even though the data sets might not belong to the same domain as that of the extractor system that is being built, allows valid extraction quality comparisons across systems and methods.

Future research needs

The next logical step in the stream of this research is building appropriate learning models and training machines that utilize our method's high-precision entity extraction. Because the entities extracted using the method discussed in this paper have very high precision and are run against a decent size of document set, they will make a very good training set for ML algorithms. Different types of ML methods could be used, such as Neural Networks [23], Support Vector Machines [24], Decision Trees, Bayesian Networks, Automated Rule Construction, Linear and Extended Linear Models, Clustering or Ensemble Learning (combination of a variety of ML methods), etc. Exploring different techniques for which one of them will give the best results, and whether different techniques can capture different (or similar) sets of previously undiscovered entities, can be interesting future research projects. Utilizing entities extracted in this paper acting as a highly reliable training set for ML will be a step toward building higher quality entity extractors. Measuring how much recall and, consequently, the F measure can be improved through different machine learning can be another aim for future research.

Finally, our proposed algorithm can be applicable to Germanic and Romance languages because they have similar features to English. An interesting research direction would be to investigate if a similar method could be developed for languages with features that are different from English, such as Asian and Semitic languages.

APPENDIX

TABLE 1
PRECISION TESTING OF EXTRACTION SYSTEMS

Test results	**Entity**	
	Person	**Organization**
Entities found (by system)	45,487	115,967
Unique entities found	14,851	22,820

Testing (random) sample size	577	585
Correct entities identified (by system)	570	570
NER precision	98.4%	97.5%

REFERENCES

Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36(4):1165–1188

Corbett P, Copestake A (2008) Cascaded classifiers for confidence-based chemical named entity recognition. In: BioNLP '08: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing., pp 54–62

Desmet B, Hoste V (2013) Fine-grained Dutch named entity recognition. Lang Resour Eval 48(2):307–343

Dozier C, Kondadadi R, Light M, Vachher A, Veeramachaneni S, Wudali R (2010) Named entity recognition and resolution in legal text. Springer, Berlin

Ekbal A, Saha S, Singh D (2012) Active machine learning technique for named entity recognition. In: ICACCI '12 Proceedings of the International Conference on Advances in Computing, Communications and Informatics., pp 180–186

Gotoh Y, Renals S (2000) Information extraction from broadcast news. Philos Trans 359(1769):1295–1310

Habib MS, Kalita J (2010) Scalable biomedical named entity recognition: investigation of a database-supported SVM approach. Int J Bioinforma Res Appl 6(2):191–208

Hakenberg J, Leaman R, Ha VN, Jonnalagadda S, Sullivan R, Miller C, Tari L, Baral C, Gonzalez G (2010) Efficient extraction of protein-protein interactions from full-text articles. IEEE/ACM Trans Comput Biol Bioinform 7(3):481–494

Hossain MS, Butler P, Boedihardjo AP, Ramakrishnan N (2012) Storytelling in entity networks to support intelligence analysts. In: KDD '12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining., pp 1375–1383

Isozaki H, Kazawa H (2002) Efficient support vectors for named entity recognition. In: COLING '02: Proceedings of the 19th international conference on Computational linguistics., pp 1–7

Kim GH, Trimi S, Chung JH (2014) Big data applications in the government sector: a comparative analysis among leading countries. Commun ACM 57(3):78–85

Lee SM (2015) The age of quality innovation. Int J Qual Innov 1(1):1–8

Lee SM, Olson D (2010) Convergenomics: strategic innovation in the convergence era. Gower Survey, UK

Nedeau D, Sekine S (2007) A survey of named entity recognition and classification. Linguisticae Investig 30(1):3–26

Petasis G, Vichot F, Wolinski F, Paliouras G, Karkaletsis V, Spyropoulos CD (2001) Using machine learning to maintain rule-based named-entity recognition and classification systems. In: ACL '01: Proceedings of the 39th Annual Meeting of Association for Computational Linguistics., pp 426–433

Ratinov L, Roth D (2009) Design challenges and misconceptions in named entity recognition. In: CoNLL'09: Proceedings of the 13th Conference on Computational Natural Language Learning of Association for Computational Linguistics., pp 147–155

Sharda R, Delen D, Turban E (2013) Business intelligence and analytics systems for decision support. Pearson Education, New Jersey

Sutton N, Wojtulewicz L, Mehta N, Gonzalez G (2012) Automatic approaches for gene-drug interaction extraction from biomedical text: corpus and comparative evaluation. In: BioNLP '12: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing., pp 214–222
Usami Y, Cho HC, Okazaki N, Tsujii J (2011) Automatic acquisition of huge training data for bio-medical named entity recognition. In: BioNLP '11: Proceedings of Biomedical Natural Language Processing Workshop., pp 65–73
Witten IH, Frank E, Hall M (2011) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington
Zaghloul W, Lee SM, Trimi S (2009) Text classification: neural networks vs. support vector machines. Ind Manag Data Syst 109(5):708–717
Zaghouani W (2012) RENAR: a rule-based Arabic named entity recognition system. ACM Trans Asian Lang Inf Process 11(1):2:1–2:13
Zayed O, El-Beltagy S, Haggag O (2013) An approach for extracting and disambiguating Arabic persons' names using clustered dictionaries and scored patterns. In: Métais E, Meziane F, Saraee M, Sugumaran V, Vadera S (eds) Natural Language Processing and Information Systems, vol 7934, NLDB 2013. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg
Zhou G, Su J (2003) Named entity recognition using an hmm-based chunk tagger. In: ACL '02: Proceedings of the 40th Annual Meeting of Association for Computational Linguistics., pp 473–480

TRANSLATED VERSION: SPANISH

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

VERSION TRADUCIDA: ESPAÑOL

A continuación se muestra una traducción aproximada de las ideas presentadas anteriormente. Esto se hizo para dar una comprensión general de las ideas presentadas en el documento. Por favor, disculpe cualquier error gramatical y no responsabilite a los autores originales de estos errores.

INTRODUCCIÓN

En el mundo global en red de hoy en día, las personas, los bienes, los datos y el conocimiento se mueven de manera más amplia, rápida y libre a la velocidad de la luz a través de los diversos límites. Para las empresas, los gobiernos y las comunidades científicas, este nuevo entorno ha traído enormes oportunidades, así como desafíos [1]. No sólo las economías de los países están más conectadas e interdependientes, sino también las personas, los gobiernos, la política y el conocimiento [2]. Los avances en las tecnologías de la información y la comunicación (TIC) han traído un enorme aumento en la cantidad de datos creados y compartidos (big data), técnicas, tecnologías y sistemas para extraer valor de los datos. El análisis de datos se utiliza para una variedad de propósitos (negocios, seguridad y seguridad, descubrimiento científico, etc.), dominios (biología, medicina, educación, etc.) Y partes interesadas (empresas, gobiernos, científicos y consumidores) [3]. Por lo tanto, la extracción de información y valor de los datos se ha vuelto fundamental para la academia, la industria y los gobiernos.

Muchas instituciones y organizaciones están reuniendo cada vez más inteligencia mediante el procesamiento y el análisis de cantidades masivas de datos que están en formato textual y se recopilan de múltiples fuentes e idiomas. El procesamiento y el análisis de estos datos, que muy a menudo son imperfectos, incompletos y no estructurados, se han vuelto cada vez más difíciles. Así, el desarrollo de nuevas tecnologías de inteligencia y análisis (I&A) y la mejora de la calidad del procesamiento y análisis de datos se han convertido en el foco de los gobiernos, matemáticos, científicos informáticos y analistas de

datos. Las metodologías, técnicas y prácticas de extracción de valor de los datos, es decir, las I+D, han cambiado con los cambios en los tipos de datos recopilados y analizados. I&A 1.0 analizó contenido estructurado; I&A 2.0 analiza contenido no estructurado (basado en texto); y I&A 3.0, que se encuentra en las primeras etapas, se centra en el análisis de contenido móvil y basado en sensores [4].

Las áreas emergentes para el análisis de texto son (1) extracción de información (IE), que extrae automáticamente información estructurada de los documentos; (2) modelo de tema (TM): descubrir los temas principales de una colección grande y no estructurada de documentos mediante el uso de algoritmos; (3) minería de opinión: acceso, extracción, clasificación y comprensión de las opiniones expresadas en muchas fuentes, incluidas las redes sociales; el análisis de opiniones también se utiliza para la minería de opiniones; y (4) respuesta a preguntas (preguntas y respuestas) — responder preguntas fácticas (por ejemplo, Watson de IBM, Siri de Apple, Alexa de Amazon, etc.) Basado en técnicas de procesamiento estadístico del lenguaje natural (NLP), recuperación de información (IR) e interacción humano-ordenador (HCI) [4].

El propósito de este estudio es añadir a la literatura actual sobre el rendimiento del reconocimiento de entidades con nombre (NER), el bloque de construcción para los sistemas IE. En concreto, creamos un extractor de entidades para las categorías Persona y Organización, que son entidades que tienen un conjunto finito de identificadores. El método de extracción propuesto mejora (1) la calidad de extracción del sistema aumentando la precisión de la extracción de entidades como resultado de las entidades extraídas iniciales de nuestro método; nuestras entidades altamente precisas se utilizarán como semilla para la formación de otros sistemas de aprendizaje automático (ML), en todos los dominios e idiomas, eliminando así no sólo la necesidad de formación manual, sino que también ampliará la semilla (a través del aprendizaje) y, por lo tanto, seguirá aumentando la precisión y el uso de la extracción de entidades a través de dominios e idiomas; y (2) la experiencia y el uso de los expertos: los expertos dedicarán menos tiempo a la formación del sistema (como lo hacen las entidades de semillas creadas por nosotros, y gastado por ML) y también menos tiempo en la fijación de entidades defectuosas y relaciones extraídas por los sistemas, por lo que, en su lugar, dedicarán su tiempo a utilizar el resultado del sistema para tomar una mejor decisión basada en datos.

Este documento está organizado de la siguiente manera: la sección "Revisión de la literatura" presenta una revisión de la literatura relevante para los conceptos teóricos del estudio; la sección "El método propuesto" explica el método propuesto para la extracción de entidades; la sección "Discusión" proporciona y analiza los resultados del estudio; y la sección "Conclusiones" concluye el estudio resumiendo las contribuciones, limitaciones y necesidades futuras de investigación del estudio.

CONCLUSIÓN

Contribuciones

El mayor beneficio de nuestro método propuesto es que crea extractores de entidades de alta precisión para las categorías Persona y Organización. Estas entidades extraídas crearán un conjunto de formación altamente fiable para algoritmos de aprendizaje automático para aprender las reglas de extracción para las dos categorías y, por lo tanto, mejorar la calidad de extracción (medida F) de los dos extractores. El número de documentos necesarios para la formación suele ser muy grande y requeriría cientos de horas de cada analista involucrado. El tiempo de los especialistas no sólo es muy caro, sino también a menudo difícil de encontrar. Por lo tanto, la eliminación del tiempo de los especialistas para capacitar a los sistemas no sólo se traduce en enormes ahorros de costos y también aumenta la eficiencia de los especialistas a medida que dedican su tiempo a realizar su trabajo mediante el uso e interpretación del resultado de los sistemas NER, en lugar de capacitar, comprobar los resultados y corregir errores. La mejora en la precisión en la extracción de entidades mejora la precisión de la extracción de relaciones de entidad, minimizando así el tiempo empleado de los usuarios del sistema en la búsqueda a través de gráficos y la fijación de relaciones defectuosas.

Otro beneficio importante de nuestro sistema es su generalización: (1) No hay necesidad de conocimiento en ningún dominio específico, ya que las dos categorías a las que aplicamos nuestro sistema son de carácter general y ampliamente utilizadas en muchos ámbitos del conocimiento. Identificar a una

persona u organización es similar en todos los dominios. Por lo tanto, no hay necesidad de expertos en dominios para construir su propio sistema de extracción. Después de una inversión inicial en la construcción de un conjunto de buenos extractores, los extractores pueden llegar a ser estables con el tiempo a través de aplicaciones y ser útiles en múltiples dominios. El método se puede perfeccionar aún más a través de los comentarios de los usuarios. (2) El método propuesto para extraer el tipo de Persona y Organización de las entidades puede aplicarse efectivamente a cualquier otra categoría de entidades que tengan un conjunto finito de subgrafios o identificadores. Si la categoría de destino no es directamente aplicable, es posible generar listas de entidades de inicialización iniciales de alta precisión similares (siempre que el número de elementos de inicialización sea lo suficientemente grande), que se puede utilizar para entrenar sistemas de aprendizaje automático para aprender las reglas de extracción. La forma más fácil de obtener estas listas de entidades de inicialización es mediante el uso de diccionarios especializados creados específicamente con un subconjunto del conjunto de entidades conocido. Por ejemplo, una categoría como Locations podría beneficiarse del uso de un diccionario especializado o un boletín. Se podría extraer una lista de inicializaciones de ubicaciones del conjunto de documentos; a continuación, la característica que rodea cada inicialización podría ser utilizada por un algoritmo de ML para generar reglas de extracción para buscar entidades. (3) En nuestro método propuesto, no es necesario etiquetar manualmente conjunto de documentos en cualquier tipo o ML en la primera etapa, aumentando la generalizabilidad de las aplicaciones (no sólo dominio) del método en muchas áreas, sin necesidad de modificaciones importantes.

Por último, nuestro método propuesto también podría utilizarse fácilmente en diferentes idiomas que tienen características similares al inglés, especialmente los idiomas germánicos (como alemán, holandés, danés, noruego, etc.) Y lenguas romances (como español, francés e italiano). Reemplazar la lista en inglés de identificadores para cada categoría por la lista equivalente del idioma de destino produciría resultados similares con poca o ninguna complicación.

Limitaciones

El objetivo principal de nuestro estudio fue mejorar la precisión de la extracción de entidades. La calidad de nuestro sistema de extracción, sin embargo, depende de la calidad de las listas de determiner. Si las categorías no tienen un número finito de "miembros", nuestro método no lograría resultados similares de alta precisión. La creación de tales listas requiere investigación y tiempo y puede variar de un idioma a otro. Este método podría ser un desafío para sistemas de uso intensivo de datos muy grandes. Sin embargo, no sería una tarea muy difícil tomar una lista en inglés de determinantes y encontrar la lista equivalente en otros idiomas.

Otra debilidad de nuestro método es que no es aplicable a todas las categorías posibles de extracción de entidades. Una categoría como Locations podría beneficiarse en gran medida del uso de un diccionario especializado o un boletín (referencia). Como se ha explicado anteriormente, se podía extraer una lista de inicializaciones de ubicaciones del conjunto de documentos y, a continuación, los algoritmos de ML podrían usar la característica que rodea cada inicialización para generar reglas de extracción o para buscar entidades similares. Estas características podrían basarse en etiquetas de parte del habla (POS), construcciones gramaticales, semántica y muchas otras características posibles. Los esfuerzos para crear estas listas aumentarán la capacidad de utilizarlas en todos los dominios de aplicación, con pocos o ningún cambio.

La última limitación en el uso de nuestro método propuesto proviene del hecho de que, dado que tiene por objeto maximizar la precisión, podría reducir la retirada como resultado indirecto. Medir la retirada en un sistema que procesa un gran número de documentos siempre es un desafío, y por lo tanto, la recuperación en tales sistemas se suele estimar. Dicha deficiencia podría superarse midiendo con precisión la recuperación basada en conjuntos de datos como conll03 [22] u otros conjuntos de datos similares. Esta evaluación de recuperación (no solo estimación) en un conjunto de datos común, aunque los conjuntos de datos podrían no pertenecer al mismo dominio que el del sistema extractor que se está construyendo, permite comparaciones de calidad de extracción válidas entre sistemas y métodos.

Necesidades de investigación futuras

El siguiente paso lógico en el flujo de esta investigación es la construcción de modelos de aprendizaje adecuados y máquinas de entrenamiento que utilizan la extracción de entidades de alta precisión de nuestro

método. Dado que las entidades extraídas mediante el método descrito en este documento tienen una precisión muy alta y se ejecutan con un tamaño decente del conjunto de documentos, crearán un muy buen conjunto de entrenamiento para algoritmos de APRENDIZAJE automático. Se podrían utilizar diferentes tipos de métodos de ML, como redes neuronales [23], máquinas vectoriales de soporte [24], árboles de decisión, redes bayesianas, construcción de reglas automatizadas, modelos lineales y lineales extendidos, clustering o aprendizaje en conjunto (combinación de una variedad de métodos de ML), etc. Explorar diferentes técnicas para las que una de ellas dará los mejores resultados, y si diferentes técnicas pueden capturar diferentes (o similares) conjuntos de entidades previamente no descubiertas, puede ser interesante proyectos de investigación futuros. El uso de entidades extraídas en este documento que actúa como un conjunto de capacitación altamente confiable para ML será un paso hacia la construcción de extractores de entidades de mayor calidad. Medir la cantidad de memoria y, en consecuencia, la medida F se puede mejorar a través de diferentes aprendizaje automático puede ser otro objetivo para la investigación futura.

Por último, nuestro algoritmo propuesto puede ser aplicable a los idiomas germánico y románico porque tienen características similares al inglés. Una dirección de investigación interesante sería investigar si se pudiera desarrollar un método similar para idiomas con características diferentes del inglés, como los idiomas asiáticos y semíticos.

TRANSLATED VERSION: FRENCH

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

VERSION TRADUITE: FRANÇAIS

Voici une traduction approximative des idées présentées ci-dessus. Cela a été fait pour donner une compréhension générale des idées présentées dans le document. Veuillez excuser toutes les erreurs grammaticales et ne pas tenir les auteurs originaux responsables de ces erreurs.

INTRODUCTION

Dans le monde mondial en réseau d'aujourd'hui, les personnes, les biens, les données et les connaissances se déplacent plus largement, rapidement et librement dans la vitesse de la lumière à travers les différentes frontières. Pour les entreprises, les gouvernements et les communautés scientifiques, ce nouvel environnement a apporté d'énormes possibilités, ainsi que des défis [1]. Non seulement les économies des pays sont-elles plus connectées et interdépendantes, mais aussi les gens, les gouvernements, la politique et le savoir [2]. Les progrès des technologies de l'information et de la communication (TIC) ont entraîné une augmentation considérable de la quantité de données créées et partagées (Big Data), techniques, technologies et systèmes pour extraire de la valeur des données. L'analyse des données est utilisée à diverses fins (affaires, sécurité et sûreté, découverte scientifique, etc.), domaines (biologie, médecine, éducation, etc.) Et parties prenantes (entreprises, gouvernements, scientifiques et consommateurs) [3]. Par conséquent, l'extraction de l'information et de la valeur des données est devenue essentielle pour le milieu universitaire, l'industrie et les gouvernements.

De nombreuses institutions et organisations recueillent de plus en plus de renseignements en traitant et en analysant des quantités massives de données qui sont en format textuel et recueillies à partir de plusieurs sources et langues. Le traitement et l'analyse de ces données, qui sont très souvent imparfaites, incomplètes et non structurées, sont devenus de plus en plus difficiles. Ainsi, le développement de nouvelles technologies d'intelligence et d'analyse (I&A) et l'amélioration de la qualité du traitement et de l'analyse des données sont devenus le centre des gouvernements, des mathématiciens, des informaticiens et des analystes de données. Les méthodes, les techniques et les pratiques d'extraction de la valeur à partir des données, c'est-à-dire les I&A, ont changé avec les changements dans les types de données recueillies et

analysées. I&A 1.0 a analysé le contenu structuré; I&A 2.0 analyse le contenu non structuré (basé sur le texte); et I&A 3.0, qui en est aux premiers stades, se concentre sur l'analyse du contenu mobile et basé sur les capteurs [4].

Les domaines émergents pour l'analyse de texte sont (1) extraction d'informations (IE) — extraction automatique d'informations structurées à partir de documents; (2) modèle de sujet (TM) — découvrir les principaux thèmes d'une vaste collection de documents non structurée à l'aide d'algorithmes; (3) l'extraction d'opinion — l'accès, l'extraction, la classification et la compréhension des opinions exprimées dans de nombreuses sources, y compris les réseaux sociaux; l'analyse des sentiments est également utilisée pour l'extraction d'opinions; et (4) répondre aux questions (Q&A) — répondre à des questions factuelles (p. Ex., Watson d'ibm, Siri d'Apple, Alexa d'Amazon, etc.) Basées sur des techniques de traitement statistique du langage naturel (NLP), de récupération d'informations (IR) et d'interaction homme-ordinateur (HCI) [4].

Le but de cette étude est d'ajouter à la littérature actuelle sur la performance de la reconnaissance des entités nommées (NER), le bloc de construction pour les systèmes IE. Plus précisément, nous construisons un extracteur d'entité pour les catégories Personne et Organisation, qui sont des entités qui ont un ensemble fini d'identificateurs. La méthode d'extraction proposée améliore (1) la qualité d'extraction du système en augmentant la précision de l'extraction de l'entité à la suite des entités extraites initiales de notre méthode; nos entités très précises seront utilisées comme graine pour la formation d'autres systèmes d'apprentissage automatique (ML), à travers les domaines et les langues, éliminant ainsi non seulement le besoin de formation manuelle, mais aussi l'expansion de la semence (par l'apprentissage) et continuera donc à augmenter la précision et l'utilisation de l'extraction d'entités à travers les domaines et les langues; et (2) l'expérience et l'utilisation des experts : les experts passeront moins de temps pour la formation du système (comme le font les entités de semences créées par nous, et dépensées par ML) et aussi moins de temps pour réparer les entités et les relations défectueuses extraites par les systèmes, donc, au lieu de cela, passeront leur temps à utiliser les résultats du système pour prendre une meilleure décision axée sur les données.

Cet article est organisé comme suit : la section « Revue de la littérature » présente un examen de la littérature pertinente pour les concepts théoriques de l'étude; la section « La méthode proposée » explique la méthode proposée pour l'extraction des entités; la section « Discussion » fournit et discute des résultats de l'étude; et la section « Conclusions » conclut l'étude en résumant les contributions, les limites et les besoins futurs de recherche de l'étude.

CONCLUSION

Contributions

Le plus grand avantage de notre méthode proposée est qu'elle crée des extracteurs d'entités de haute précision pour les catégories Personne et Organisation. Ces entités extraites feront un ensemble de formation très fiable pour les algorithmes d'apprentissage automatique pour apprendre les règles d'extraction pour les deux catégories, et ainsi améliorer la qualité d'extraction (mesure F) des deux extracteurs. Le nombre de documents nécessaires à la formation est généralement très important et nécessiterait des centaines d'heures de la part de chaque analyste concerné. Le temps des spécialistes est non seulement très coûteux, mais aussi souvent difficile à trouver. Ainsi, l'élimination du temps des spécialistes pour la formation des systèmes entraîne non seulement d'énormes économies de coûts, mais augmente également l'efficacité des spécialistes à mesure qu'ils passent leur temps à accomplir leur travail en utilisant et en interprétant les résultats des systèmes NER, au lieu de former, vérifier les résultats et corriger les erreurs. L'amélioration de la précision dans l'extraction des entités améliore la précision de l'extraction des relations d'entité, minimisant ainsi le temps passé par les utilisateurs du système à rechercher des graphiques et à corriger les relations défectueuses.

Un autre avantage important de notre système est sa généralisabilité: (1) Il n'y a pas besoin de connaissances dans un domaine spécifique, car les deux catégories auxquelle nous avons appliqué notre système sont à la fois de nature générale et largement utilisées dans de nombreux domaines de connaissances. L'identification d'une personne ou d'une organisation est similaire dans tous les domaines.

Ainsi, il n'est pas nécessaire que les experts du domaine construisent leur propre système d'extraction. Après un investissement initial dans la construction d'un ensemble de bons extracteurs, les extracteurs peuvent devenir stables au fil du temps grâce à des applications et être utiles dans plusieurs domaines. La méthode peut être affinée grâce à la rétroaction des utilisateurs. (2) La méthode proposée pour extraire le type de personne et d'organisation des entités peut être appliquée efficacement à toute autre catégorie d'entités qui ont un ensemble fini de sous-types ou d'identificateurs. Si la catégorie cible n'est pas directement applicable, il est possible de générer des listes initiales d'entités de semences de haute précision similaires (tant que le nombre d'éléments de semences est assez grand), qui peuvent être utilisés pour former des systèmes d'apprentissage automatique pour apprendre les règles d'extraction. La façon la plus simple d'obtenir de telles listes d'entités de semences est l'utilisation de dictionnaires spécialisés spécialement construits avec un sous-ensemble de l'ensemble d'entités connues. Par exemple, une catégorie comme Locations pourrait bénéficier de l'utilisation d'un dictionnaire spécialisé ou d'un gazetter. Une liste de semences des emplacements pourrait être extraite de l'ensemble de documents; alors la fonctionnalité entourant chaque graine pourrait être utilisée par un algorithme ML pour générer des règles d'extraction pour trouver des entités. (3) Dans notre méthode proposée, il n'est pas nécessaire d'étiqueter manuellement l'ensemble de documents de formation en aucune sorte ou ML dans la première étape, ce qui augmente la généralisation des applications (et pas seulement le domaine) de la méthode dans de nombreux domaines, sans la nécessité de modifications majeures.

Enfin, notre méthode proposée pourrait également être facilement utilisée dans différentes langues qui ont des caractéristiques similaires à l'anglais, en particulier les langues germaniques (comme l'allemand, le néerlandais, le danois, le norvégien, etc.) Et les langues romanes (comme l'espagnol, le Français et l'italien). Le remplacement de la liste anglaise des identificateurs de chaque catégorie par la liste équivalente de la langue cible donnerait des résultats similaires avec peu ou pas de complications.

Limitations

L'objectif principal de notre étude était d'améliorer la précision de l'extraction des entités. La qualité de notre système d'extraction, cependant, dépend de la qualité des listes de déterminants. Si les catégories n'ont pas un nombre limité de « membres », notre méthode n'obtiendrait pas des résultats similaires de haute précision. La création de telles listes nécessite des recherches et du temps et peut varier d'une langue à l'autre. Cette méthode pourrait être un défi pour les très grands systèmes à forte intensité de données. Il ne serait toutefois pas très difficile de prendre une liste anglaise de déterminants et de trouver la liste équivalente dans d'autres langues.

Une autre faiblesse de notre méthode est qu'elle ne s'applique pas à toutes les catégories d'extraction d'entités possibles. Une catégorie comme Locations pourrait fortement bénéficier de l'utilisation d'un dictionnaire spécialisé ou d'un gazetter (référence). Comme nous l'avons vu précédemment, une liste de semences d'emplacements pourrait être extraite de l'ensemble de documents, puis, la fonctionnalité entourant chaque graine pourrait être utilisée par les algorithmes ML pour générer des règles d'extraction ou pour trouver des entités similaires. Ces fonctionnalités peuvent être basées sur une partie des balises de la parole (POS), des constructions grammaticales, de la sémantique et de nombreuses autres fonctionnalités possibles. Les efforts pour créer de telles listes augmenteront la capacité de les utiliser dans tous les domaines d'application, avec peu ou pas de changements.

La dernière limite à l'utilisation de notre méthode proposée vient du fait que, parce qu'elle vise à maximiser la précision, elle pourrait réduire le rappel en conséquence indirecte. La mesure du rappel dans un système qui traite un très grand nombre de documents est toujours un défi, et par conséquent, le rappel dans de tels systèmes est habituellement estimé. Une telle lacune pourrait être surmontée en mesurant avec précision le rappel en fonction d'ensembles de données tels que conll03 [22] ou d'autres ensembles de données similaires. Cette évaluation de rappel (et pas seulement l'estimation) sur un ensemble de données commun, même si les ensembles de données peuvent ne pas appartenir au même domaine que celui du système d'extraction en cours de construction, permet des comparaisons valides de qualité d'extraction entre les systèmes et les méthodes.

Besoins futurs en recherche

La prochaine étape logique dans le flux de cette recherche est la construction de modèles d'apprentissage appropriés et des machines de formation qui utilisent l'extraction de l'entité de haute précision de notre méthode. Parce que les entités extraites en utilisant la méthode discutée dans ce document ont une très grande précision et sont exécutés contre une taille décente de l'ensemble de documents, ils feront un ensemble de formation très bon pour les algorithmes ML. Différents types de méthodes ML pourraient être utilisés, tels que les réseaux neuronaux [23], les machines vectorielles de support [24], les arbres de décision, les réseaux bayésiens, la construction automatisée de règles, les modèles linéaires et étendus, le regroupement ou l'apprentissage d'ensemble (combinaison d'une variété de méthodes ML), etc. Explorer différentes techniques pour lesquelles l'une d'entre elles donnera les meilleurs résultats, et si différentes techniques peuvent capturer différents (ou similaires) ensembles d'entités précédemment inconnues, peut être intéressant projets de recherche futurs. L'utilisation d'entités extraites dans ce document agissant comme un ensemble de formation hautement fiable pour ML sera une étape vers la construction d'extracteurs d'entités de meilleure qualité. Mesurer la quantité de rappel et, par conséquent, la mesure F peut être améliorée grâce à différents apprentissages automatiques peut être un autre objectif pour la recherche future.

Enfin, notre algorithme proposé peut s'appliquer aux langues germanique et romane car elles ont des caractéristiques similaires à l'anglais. Une direction de recherche intéressante serait d'étudier si une méthode similaire pourrait être développée pour les langues avec des caractéristiques qui sont différentes de l'anglais, comme les langues asiatiques et sémitiques.

TRANSLATED VERSION: GERMAN

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

ÜBERSETZTE VERSION: DEUTSCH

Hier ist eine ungefähre Übersetzung der oben vorgestellten Ideen. Dies wurde getan, um ein allgemeines Verständnis der in dem Dokument vorgestellten Ideen zu vermitteln. Bitte entschuldigen Sie alle grammatikalischen Fehler und machen Sie die ursprünglichen Autoren nicht für diese Fehler verantwortlich.

EINLEITUNG

In der heutigen vernetzten globalen Welt bewegen sich Menschen, Waren, Daten und Wissen breiter, schneller und freier in der Lichtgeschwindigkeit über die verschiedenen Grenzen hinweg. Für Unternehmen, Regierungen und wissenschaftliche Gemeinschaften hat dieses neue Umfeld enorme Chancen und Herausforderungen mit sich gebracht [1]. Die Volkswirtschaften der Länder sind nicht nur stärker vernetzt und voneinander abhängig, sondern auch die Menschen, Regierungen, Politik und Wissen [2]. Die Fortschritte in der Informations- und Kommunikationstechnologie (IKT) haben zu einer enormen Zunahme der Menge an erzeugten und geteilten Daten (Big Data), Techniken, Technologien und Systemen geführt, um den Daten Einen Mehrwert zu ziehen. Datenanalysen werden für eine Vielzahl von Zwecken (Geschäft, Sicherheit und Sicherheit, wissenschaftliche Entdeckung usw.), Bereichen (Biologie, Medizin, Bildung usw.) Und Stakeholdern (Unternehmen, Regierungen, Wissenschaftler und Verbraucher) verwendet [3]. Daher ist das Extrahieren von Informationen und Nutzen aus Daten für die Wissenschaft, die Industrie und Die Regierungen von entscheidender Bedeutung geworden.

Viele Institutionen und Organisationen sammeln zunehmend Informationen, indem sie riesige Datenmengen im Textformat verarbeiten und analysieren, die aus mehreren Quellen und Sprachen gesammelt werden. Die Verarbeitung und Analyse solcher Daten, die sehr oft unvollkommen, unvollständig und unstrukturiert sind, wird immer schwieriger. So stehen die Entwicklung neuer Intelligenz- und

Analysetechnologien (I&A) und die Verbesserung der Qualität der Datenverarbeitung und -analyse im Fokus von Regierungen, Mathematikern, Informatikern und Datenanalysten. Methoden, Techniken und Praktiken zum Extrahieren von Wert aus Daten, d. H. I&A, haben sich mit den Änderungen in den gesammelten und analysierten Datentypen geändert. I&A 1.0 analysierte strukturierte Inhalte; I&A 2.0 analysiert unstrukturierte (textbasierte) Inhalte; und I&A 3.0, das sich in der Anfangsphase befindet, konzentriert sich auf die mobile und sensorbasierte Content-Analyse [4].

Die neu entstehenden Bereiche für Textanalysen sind (1) Die Extraktion von Informationen (IE) – automatisches Extrahieren strukturierter Informationen aus Dokumenten; (2) Themenmodell (TM) – Ermittlung der Hauptthemen in einer großen und unstrukturierten Sammlung von Dokumenten mithilfe von Algorithmen; (3) Meinungsabbau – Zugang, Auszug, Klassifizierung und Verständnis der in vielen Quellen, einschließlich sozialer Netzwerke, geäußerten Meinungen; Stimmungsanalyse wird auch für Denmeinungsbergbau verwendet; und (4) Fragebeantwortung (Q&A) – Beantwortung sachlicher Fragen (z. B. WATSON von IBM, Siri von Apple, Amazons Alexa usw.) Basierend auf Techniken aus der statistischen Verarbeitung natürlicher Sprachen (NLP), dem Informationsabruf (IR) und der Mensch-Computer-Interaktion (HCI) [4].

Der Zweck dieser Studie besteht darin, die aktuelle Literatur über die Leistung der benannten Entitätserkennung (NER), den Baustein für IE-Systeme, zu ergänzen. Insbesondere erstellen wir einen Entitätsextraktor für die Kategorien Person und Organisation, d. H. Entitäten, die über einen endlichen Satz von Bezeichnern verfügen. Die vorgeschlagene Extraktionsmethode verbessert (1) die Extraktionsqualität des Systems, indem die Genauigkeit der Entitätsextraktion infolge der anfänglichen extrahierten Einheiten aus unserer Methode erhöht wird; unsere hochpräzisen Einheiten werden als Saatgut für die Ausbildung anderer Machine Learning (ML)-Systeme, über Domänen und Sprachen hinweg, verwendet werden, wodurch nicht nur die Notwendigkeit manueller Schulungen entfällt, sondern auch das Saatgut (durch Lernen) erweitert wird und daher die Präzision und Verwendung der Entitätsextraktion über Domänen und Sprachen hinweg weiter erhöht wird; und (2) die Erfahrung und nutzung der Experten: Experten werden weniger Zeit für die Schulung des Systems aufwenden (wie dies von den von uns geschaffenen und von ML erstellten Saatguteinheiten geschieht) und auch weniger Zeit für die Behebung fehlerhafter Entitäten und Beziehungen, die von den Systemen extrahiert werden, und damit stattdessen ihre Zeit damit verbringen, das Systemergebnis zu nutzen, um eine bessere datengesteuerte Entscheidung zu treffen.

Dieses Papier ist wie folgt organisiert: die Rubrik "Literaturrezension" präsentiert einen Überblick über relevante Literatur für die theoretischen Konzepte der Studie; im Abschnitt "Die vorgeschlagene Methode" wird die vorgeschlagene Methode für die Gewinnung von Unternehmen erläutert; im Abschnitt "Diskussion" werden die Studienergebnisse erläutert und erörtert; und der Abschnitt "Schlussfolgerungen" schließt die Studie mit einer Zusammenfassung der Beiträge, Einschränkungen und des künftigen Forschungsbedarfs der Studie ab.

SCHLUSSFOLGERUNG

Beiträge

Der größte Vorteil unserer vorgeschlagenen Methode besteht darin, dass hochpräzise Entitätsextraktoren für die Kategorien Person und Organisation erstellt werden. Diese extrahierten Entitäten bilden ein sehr zuverlässiges Trainingsset für Machine Learning-Algorithmen, um die Extraktionsregeln für die beiden Kategorien zu erlernen und so die Extraktionsqualität (F-Messung) aller beiden Extraktoren zu verbessern. Die Anzahl der Dokumente, die für die Schulung benötigt werden, ist in der Regel sehr groß und würde Hunderte von Stunden von jedem beteiligten Analysten erfordern. Die Zeit der Spezialisten ist nicht nur sehr teuer, sondern oft auch schwer zu finden. Die Eliminierung der Zeit für die Schulung von Systemen führt nicht nur zu enormen Kosteneinsparungen und erhöht auch die Effizienz von Spezialisten, da sie ihre Arbeit damit verbringen, das Ergebnis von NER-Systemen zu nutzen und zu interpretieren, anstatt Schulungen durchzuführen, Ergebnisse zu überprüfen und Fehler zu korrigieren. Die Verbesserung der Genauigkeit bei der Extraktion von Entitäten verbessert die Genauigkeit der Extraktion von

Entitätsbeziehungen und minimiert so die Zeit der Systembenutzer für die Suche nach Graphen und die Behebung fehlerhafter Beziehungen.

Ein weiterer wichtiger Vorteil unseres Systems ist seine Verallgemeinerbarkeit: (1) Es besteht kein Bedarf an Wissen in einem bestimmten Bereich, da die beiden Kategorien, auf die wir unser System angewendet haben, sowohl allgemeiner Natur sind als auch in vielen Wissensbereichen weit verbreitet sind. Das Identifizieren einer Person oder Organisation ist domänenübergreifend ähnlich. Daher ist es nicht notwendig, dass Domain-Experten ihr eigenes Extraktionssystem bauen. Nach einer anfänglichen Investition in den Aufbau einer Reihe von guten Extraktoren können die Extraktoren im Laufe der Zeit durch Anwendungen stabil werden und in mehreren Domänen nützlich sein. Die Methode kann durch das Feedback der Benutzer weiter verfeinert werden. (2) Die vorgeschlagene Methode zum Extrahieren des Personen- und Organisationstyps von Entitäten kann effektiv auf jede andere Kategorie von Entitäten angewendet werden, die über einen endlichen Satz von Untertypen oder Bezeichnern verfügen. Wenn die Zielkategorie nicht direkt anwendbar ist, ist es möglich, ähnliche anfängliche hochpräzise Seed-Entitätslisten zu generieren (solange die Anzahl der Seed-Elemente groß genug ist), die verwendet werden können, um maschinelle Lernsysteme zu trainieren, um die Extraktionsregeln zu erlernen. Der einfachste Weg, solche Seed-Entitätslisten zu erhalten, ist die Verwendung spezieller Wörterbücher, die speziell mit einer Teilmenge des bekannten Entitätssatzes erstellt wurden. Beispielsweise könnte eine Kategorie wie Locations von der Verwendung eines spezialisierten Wörterbuchs oder eines Gazetteers profitieren. Eine Seed-Liste der Speicherorte könnte aus dem Dokumentsatz extrahiert werden. Dann könnte das Feature, das jeden Seed umgibt, von einem ML-Algorithmus verwendet werden, um Extraktionsregeln für die Suche nach Entitäten zu generieren. (3) In unserer vorgeschlagenen Methode ist es nicht erforderlich, in der ersten Phase manuell markierte Schulungssets von Dokumenten in irgendeiner Art oder ML zu erstellen, wodurch die Verallgemeinerbarkeit von Anwendungen (nicht nur domäne) der Methode in vielen Bereichen erhöht wird, ohne dass größere Änderungen erforderlich sind.

Schließlich könnte unsere vorgeschlagene Methode auch problemlos in verschiedenen Sprachen verwendet werden, die ähnliche Merkmale wie Englisch aufweisen, insbesondere in germanischen Sprachen (z. B. Deutsch, Niederländisch, Dänisch, Norwegisch usw.) Und romanischen Sprachen (z. B. Spanisch, Französisch und Italienisch). Das Ersetzen der englischen Liste der Bezeichner für jede Kategorie durch die entsprechende Liste aus der Zielsprache würde zu ähnlichen Ergebnissen führen, ohne dass es zu Komplikationen kommt.

Einschränkungen

Das Hauptziel unserer Studie war es, die Präzision der Entitätsextraktion zu verbessern. Die Qualität unseres Absaugsystems hängt jedoch von der Qualität der Bestimmendenlisten ab. Wenn die Kategorien keine endliche Anzahl von "Mitgliedern" haben, würde unsere Methode keine ähnlichen hochpräzisen Ergebnisse erzielen. Die Erstellung solcher Listen erfordert Forschung und Zeit und kann von Sprache zu Sprache variieren. Diese Methode könnte eine Herausforderung für sehr große datenintensive Systeme sein. Es wäre jedoch keine sehr schwierige Aufgabe, eine englische Liste von Bestimmenden zu erstellen und die entsprechende Liste in anderen Sprachen zu finden.

Eine weitere Schwachstelle unserer Methode ist, dass sie nicht auf alle möglichen Entitätsextraktionskategorien anwendbar ist. Eine Kategorie wie Locations könnte von der Verwendung eines spezialisierten Wörterbuchs oder eines Gazetteers (Referenz) sehr profitieren. Wie bereits erwähnt, könnte eine Seed-Liste von Speicherorten aus dem Dokumentsatz extrahiert werden, und dann könnte das Feature, das jeden Seed umgibt, von ML-Algorithmen verwendet werden, um Extraktionsregeln zu generieren oder ähnliche Entitäten zu finden. Solche Funktionen können auf einem Teil von Sprach-Tags (POS), grammatikalischen Konstrukten, Semantik und vielen anderen möglichen Features basieren. Die Bemühungen, solche Listen zu erstellen, werden die Fähigkeit erhöhen, sie über Anwendungsdomänen hinweg zu nutzen, ohne dass sich Änderungen ändern.

Die letzte Einschränkung bei der Anwendung unserer vorgeschlagenen Methode ergibt sich aus der Tatsache, dass sie, da sie auf die Maximierung der Genauigkeit abzielt, den Rückruf als indirektes Ergebnis senken könnte. Die Messung des Rückrufs in einem System, das eine sehr große Anzahl von Dokumenten verarbeitet, ist immer eine Herausforderung, und daher wird der Rückruf in solchen Systemen in der Regel

geschätzt. Ein solcher Mangel könnte durch eine genaue Messung des Rückrufs auf der Grundlage von Datensätzen wie conll03 [22] oder anderen ähnlichen Datensätzen überwunden werden. Diese Rückrufauswertung (nicht nur Dieschätzung) für einen gemeinsamen Datensatz, auch wenn die Datensätze möglicherweise nicht zur gleichen Domäne wie das zu entwickelnde Extraktorsystem gehören, ermöglicht valide Extraktionsqualitätsvergleiche zwischen Systemen und Methoden.

Zukünftiger Forschungsbedarf

Der nächste logische Schritt im Strom dieser Forschung ist die Erstellung geeigneter Lernmodelle und Trainingsmaschinen, die die hochpräzise Entitätsextraktion unserer Methode nutzen. Da die Entitäten, die mit der in diesem Dokument beschriebenen Methode extrahiert wurden, eine sehr hohe Genauigkeit aufweisen und gegen eine anständige Größe des Dokumentsatzes ausgeführt werden, werden sie einen sehr guten Trainingssatz für ML-Algorithmen bilden. Es könnten verschiedene Arten von ML-Methoden verwendet werden, wie z. B. Neuronale Netzwerke [23], Support Vector Machines [24], Entscheidungsbäume, Bayesian Networks, Automated Rule Construction, Linear and Extended Linear Models, Clustering oder Ensemble Learning (Kombination einer Vielzahl von ML-Methoden) usw. Die Untersuchung verschiedener Techniken, für die eine von ihnen die besten Ergebnisse liefert, und ob verschiedene Techniken verschiedene (oder ähnliche) Gruppen von zuvor unentdeckten Entitäten erfassen können, kann in zukünftigen Forschungsprojekten interessant sein. Die Verwendung von Entitäten, die in diesem Dokument extrahiert werden und als sehr zuverlässiges Schulungsset für ML fungieren, wird ein Schritt zum Aufbau hochwertigerer Entitätsextraktoren sein. Die Messung, wie viel Rückruf und damit die F-Maßnahme durch unterschiedliches maschinelles Lernen verbessert werden können, kann ein weiteres Ziel für die zukünftige Forschung sein.

Schließlich kann unser vorgeschlagener Algorithmus auf germanische und romanische Sprachen anwendbar sein, da sie ähnliche Funktionen wie Englisch aufweisen. Eine interessante Forschungsrichtung wäre zu untersuchen, ob eine ähnliche Methode für Sprachen entwickelt werden könnte, die sich von Englisch unterscheiden, wie asiatische und semitische Sprachen.

TRANSLATED VERSION: PORTUGUESE

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

VERSÃO TRADUZIDA: PORTUGUÊS

Aqui está uma tradução aproximada das ideias acima apresentadas. Isto foi feito para dar uma compreensão geral das ideias apresentadas no documento. Por favor, desculpe todos os erros gramaticais e não responsacule os autores originais responsáveis por estes erros.

INTRODUÇÃO

No mundo global em rede de hoje, as pessoas, bens, dados e conhecimento se movem mais amplamente, rapidamente e livremente na velocidade da luz através das várias fronteiras. Para empresas, governos e comunidades científicas, esse novo ambiente trouxe enormes oportunidades, além de desafios [1]. Não só as economias dos países estão mais conectadas e interdependentes, mas também as pessoas, governos, políticas e conhecimentos [2]. Os avanços na tecnologia da informação e comunicação (TIC) trouxeram um enorme aumento na quantidade de dados criados e compartilhados (big data), técnicas, tecnologias e sistemas para extrair valor dos dados. A análise de dados é usada para uma variedade de propósitos (negócios, segurança e segurança, descoberta científica, etc.), domínios (biologia, medicina, educação, etc.), e stakeholders (empresas, governos, cientistas e consumidores) [3]. Portanto, extrair informações e valor dos dados tornou-se fundamental para a academia, a indústria e os governos.

Muitas instituições e organizações estão cada vez mais coletando informações processando e analisando grandes quantidades de dados que estão em formato textual e coletados de múltiplas fontes e idiomas. O processamento e a análise desses dados, muitas vezes imperfeitos, incompletos e desestruturados, tornaram-se cada vez mais difíceis. Assim, o desenvolvimento de novas tecnologias de inteligência e análise (I&A) e a melhoria da qualidade do processamento e análise de dados tornaram-se o foco de governos, matemáticos, cientistas da computação e analistas de dados. Metodologias, técnicas e práticas de extração de valor a partir de dados, ou seja, I&A, mudaram com as mudanças nos tipos de dados coletados e analisados. I&A 1.0 analisou conteúdo estruturado; I&A 2.0 analisa conteúdo não estruturado (baseado em texto); e I&A 3.0, que está nos estágios iniciais, foca na análise de conteúdo móvel e baseada em sensores [4].

As áreas emergentes para análise de texto são (1) extração de informações (IE) — extraindo automaticamente informações estruturadas de documentos; (2) modelo de tópico (TM)— descobrindo os principais temas em uma grande e não estruturada coleção de documentos usando algoritmos; (3) mineração de opinião — acesso, extrato, classificação e compreensão das opiniões expressas em muitas fontes, incluindo redes sociais; a análise do sentimento também é utilizada para a mineração de opinião; e (4) resposta a perguntas (Q&A)— respondendo a perguntas factuais (por exemplo, Watson da IBM, Siri da Apple, Alexa da Amazon, etc.) Com base em técnicas de processamento estatístico de linguagem natural (PNL), recuperação de informações (IR) e interação homem-computador (HCI) [4].

O objetivo deste estudo é agregar à literatura atual sobre a realização do reconhecimento de entidades denominadas (NER), o bloco de construção para sistemas IE. Especificamente, construímos um extrator de entidades para categorias Pessoa e Organização, que são entidades que possuem um conjunto finito de identificadores. O método de extração proposto melhora (1) a qualidade de extração do sistema, aumentando a precisão da extração de entidades como resultado das entidades extraídas iniciais do nosso método; nossas entidades altamente precisas serão utilizadas como semente para a formação de outros sistemas de machine learning (ML), entre domínios e idiomas, eliminando assim não apenas a necessidade de treinamento manual, mas também expandirá a semente (através do aprendizado) e, portanto, continuarão a aumentar a precisão e o uso da extração de entidades entre domínios e idiomas; e (2) a experiência e o uso dos especialistas: os especialistas gastarão menos tempo para treinar o sistema (como é feito pelas entidades de sementes criadas por nós, e gastas pela ML) e também menos tempo na fixação de entidades e relacionamentos defeituosos extraídos pelos sistemas, assim, em vez disso, gastarão seu tempo usando o resultado do sistema para tomar uma melhor decisão orientada por dados.

Este artigo é organizado da seguinte forma: a seção "Revisão da Literatura" apresenta uma revisão da literatura relevante para os conceitos teóricos do estudo; a seção "O método proposto" explica o método proposto para a extração de entidades; a seção "Discussão" fornece e discute os resultados do estudo; e a seção "Conclusões" conclui o estudo resumindo as contribuições, limitações e necessidades futuras de pesquisa do estudo.

CONCLUSÃO

Contribuições

O maior benefício do nosso método proposto é que ele cria extratores de entidades de alta precisão para as categorias Pessoa e Organização. Essas entidades extraídas farão um conjunto de treinamento altamente confiável para algoritmos de aprendizagem de máquina para aprender as regras de extração para as duas categorias e, assim, melhorar a qualidade da extração (medida F) de todos os dois extratores. O número de documentos necessários para o treinamento é geralmente muito grande e exigiria centenas de horas de cada analista envolvido. O tempo dos especialistas não é apenas muito caro, mas também muitas vezes difícil de encontrar. Assim, eliminar o tempo dos especialistas para treinar sistemas não só resulta em enorme redução de custos e também aumenta a eficiência dos especialistas à medida que gastam seu tempo realizando seu trabalho usando e interpretando o resultado dos sistemas NER, em vez de treinar, verificar resultados e corrigir erros. A melhoria da precisão na extração de entidades melhora a precisão da extração de

relacionamento com entidades, minimizando assim o tempo gasto pelos usuários do sistema na busca por gráficos e fixação de relacionamentos defeituosos.

Outro benefício importante do nosso sistema é a sua generalizabilidade: (1) Não há necessidade de conhecimento em nenhum domínio específico, pois as duas categorias a que aplicamos nosso sistema são de natureza geral e amplamente utilizadas em muitos domínios do conhecimento. Identificar uma pessoa ou organização é semelhante entre os domínios. Assim, não há necessidade de especialistas em domínios construirem seu próprio sistema de extração. Depois de um investimento inicial na construção de um conjunto de bons extratores, os extratores podem se tornar estáveis ao longo do tempo através de aplicativos e serem úteis em vários domínios. O método pode ser refinado ainda mais através do feedback dos usuários. (2) O método proposto para extrair o tipo de Pessoa e Organização das entidades pode ser efetivamente aplicado a qualquer outra categoria de entidades que tenham um conjunto finito de subtipos ou identificadores. Se a categoria de destino não for diretamente aplicável, é possível gerar listas de entidades de sementes iniciais de alta precisão semelhantes (desde que o número de itens de sementes seja grande o suficiente), que podem ser usados para treinar sistemas de aprendizagem de máquina para aprender as regras de extração. A maneira mais fácil de obter tais listas de entidades de sementes é através do uso de dicionários especializados especificamente construídos com um subconjunto do conjunto de entidades conhecidas. Por exemplo, uma categoria como Locations poderia se beneficiar do uso de um dicionário especializado ou de um gazeta. Uma lista de sementes de locais poderia ser extraída do conjunto de documentos; em seguida, o recurso em torno de cada semente poderia ser usado por um algoritmo ML para gerar regras de extração para encontrar entidades. (3) Em nosso método proposto, não há necessidade de conjunto de treinamento manualmente marcado de documentos em qualquer tipo ou ML no primeiro estágio, aumentando a generalizabilidade das aplicações (não apenas domínio) do método em muitas áreas, sem a necessidade de grandes modificações.

Finalmente, nosso método proposto também poderia ser facilmente usado em diferentes idiomas que têm características semelhantes ao inglês, especialmente línguas germânicas (como, alemão, holandês, dinamarquês, norueguês, etc.) E línguas românticas (como, espanhol, francês e italiano). Substituir a lista de identificadores em inglês para cada categoria com a lista equivalente do idioma alvo produziria resultados semelhantes com pouca ou nenhuma complicação.

Limitações

O objetivo principal do nosso estudo foi melhorar a precisão da extração da entidade. A qualidade do nosso sistema de extração, no entanto, depende da qualidade das listas de determinantes. Se as categorias não tiverem um número finito de "membros", nosso método não alcançaria resultados semelhantes de alta precisão. A criação dessas listas requer pesquisa e tempo e pode variar de uma língua para outra. Este método pode ser um desafio para sistemas muito grandes de uso intensivo de dados. Não seria uma tarefa muito difícil, no entanto, pegar uma lista de determinantes em inglês e encontrar a lista equivalente em outros idiomas.

Outra fraqueza do nosso método é que ele não é aplicável a todas as categorias de extração de entidades possíveis. Uma categoria como Localizações poderia se beneficiar muito do uso de um dicionário especializado ou de um gazetteer (referência). Como discutido anteriormente, uma lista de sementes de locais poderia ser extraída do conjunto de documentos e, em seguida, o recurso em torno de cada semente poderia ser usado por algoritmos ML para gerar regras de extração ou para encontrar entidades semelhantes. Tais características poderiam ser baseadas em parte de tags de fala (PDV), construções gramaticais, semântica e muitas outras características possíveis. Os esforços para criar tais listas aumentarão a capacidade de utilizá-las entre domínios de aplicativos, com pouca ou nenhuma alteração.

A última limitação no uso do nosso método proposto vem do fato de que, por se destinar a maximizar a precisão, poderia diminuir o recall como resultado indireto. Medir o recall em um sistema que processa um número muito grande de documentos é sempre um desafio e, portanto, o recall nesses sistemas é geralmente estimado. Tal deficiência poderia ser superada medindo com precisão o recall com base em conjuntos de dados como conll03 [22] ou outros conjuntos de dados semelhantes. Esta avaliação de recall (não apenas estimativa) em um conjunto de dados comum, embora os conjuntos de dados possam não

pertencer ao mesmo domínio do sistema extrator que está sendo construído, permite comparações válidas de qualidade de extração entre sistemas e métodos.

Necessidades futuras de pesquisa

O próximo passo lógico no fluxo desta pesquisa é construir modelos de aprendizagem apropriados e máquinas de treinamento que utilizem a extração de entidades de alta precisão do nosso método. Como as entidades extraídas usando o método discutido neste artigo têm uma precisão muito alta e são executadas contra um tamanho decente de conjunto de documentos, eles farão um conjunto de treinamento muito bom para algoritmos de ML. Diferentes tipos de métodos ML poderiam ser usados, como Redes Neurais [23], Máquinas vetoriais de suporte [24], Árvores de Decisão, Redes Bayesian, Construção automatizada de regras, modelos lineares lineares e estendidos, clustering ou ensemble learning (combinação de uma variedade de métodos ML), etc. Explorar diferentes técnicas para as quais uma delas dará os melhores resultados, e se diferentes técnicas podem capturar diferentes (ou similares) conjuntos de entidades não descobertas, pode ser interessante projetos de pesquisa futuras. As entidades de utilização extraídas neste artigo atuando como um conjunto de treinamento altamente confiável para a ML será um passo para a construção de extratores de entidades de maior qualidade. Medir o quanto de recordação e, consequentemente, a medida F pode ser melhorada através de diferentes aprendizados de máquina pode ser outro objetivo para pesquisas futuras.

Finalmente, nosso algoritmo proposto pode ser aplicável às línguas germânicas e românticas porque eles têm características semelhantes ao inglês. Uma interessante direção de pesquisa seria investigar se um método semelhante poderia ser desenvolvido para línguas com características diferentes do inglês, como línguas asiáticas e semíticas.