CROSS-DOCUMENT COREFERENCE RESOLUTION MODEL BASED ON NEURAL ENTITY EMBEDDING

ALIAKBAR KESHTKARAN

UNIVERSITI TEKNOLOGI MALAYSIA

CROSS-DOCUMENT COREFERENCE RESOLUTION MODEL BASED ON NEURAL ENTITY EMBEDDING

ALIAKBAR KESHTKARAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia

JUNE 2021

DEDICATION

To my GOD, ***ALLAH***, who is always with me in every moment

To our prophet, ***Mohammad***, the messenger of truth, fraternization and kindness

To ***Mahdi*** the promised saviour, looking forward to his arrival

To my dears mother, father, sisters, and brother

To my dear and beloved wife who encouraged and supported me

To my loving daughters who gave me love

To my dears mother-, father-, and brother-in-law

And to all who supported me in my study, especially my supervisor

ACKNOWLEDGEMENT

In the name of Allah, Most Gracious, Most Merciful. I thank Allah S.W.T for granting me perseverance and strength I needed to complete this thesis.

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. First and foremost, I wish to express my sincere appreciation to my thesis supervisor, Assoc. Prof. Dr. Siti Sophiayati Yuhaniz, for encouragement, guidance, and critics. She has built and directed an environment that granted me an opportunity to learn and practice research skills, meet, and collaborate with brilliant researchers, and transfer the long journey of PhD to a great and lovely experience. I would like to thank Prof. Dr. Suhaimi Ibrahim, my first supervisor of PhD which he did his best for supporting me during his supervision.

I would particularly like to thank my parents and my wife's parents, who deserve my gratitude for their inseparable prayer, encouragement, and endless patience. Words fail me in expressing my deepest appreciation to my wife, Zahra, and my loving daughters, Ayeh and Motahareh, whose encouragement, patient, and love gave me confidence. Thesis would not have been possible without their support. Thank you.

ABSTRACT

Natural Language Processing (NLP) is a way for computers to derive, analyze, and understand the meaning of human language in a smart and useful way. NLP considers the hierarchical structure of language that enables real-world applications such as automatic text summarization, event resolution, relationship extraction, and entity recognition to be presented in a proper human-computer interaction. One of the NLP components called Coreference Resolution (CR) is to determine whether the two noun phrases in natural language are referring to the same entity. In this context, an entity can be a real person, organization, place, or others, in which the referred term of such entity is called a mention. The task of CR when extended to resolve co-referent entities across multiple documents creates the Cross-Document Coreference Resolution (CDCR) task which requires special techniques to manage and address the mention chains within documents co-referring to the same entity across different documents. Currently, there are some limitations in the existing works in which the CDCR entities by variant referencing mentions are not well identified, and the grouping process to differentiate entities with lexical similarity is not well addressed. The main objective of this research is to propose a CDCR model using neural embedding of the entities and their mentions created by the representation of words using merely the input documents. This model created vectors of mentions and entities using neural embedding of mentions, regardless of the use of any external resources such as Knowledge Bases. For an advanced grouping of entities and their mentions, an improved density-based clustering technique containing DBSCAN and H-DBSCAN clustering algorithms was employed. In addition, a prototype named CROCER was designed and developed as proof of concept to assess the model in an experimental environment. For evaluation, this model was applied to three publicly available datasets, called 'John Smith Corpus', 'WePS-2 Collection', and 'Google Wikilinks' from public open-source repositories. It measured the precision, recall, and F1 score of the model by three known scoring systems for Coreference Resolution, which are MUC, B3, and CEAF. Based on the findings, it can be concluded that the proposed model improved the F1 score of the datasets by almost 15.7%, 1.5%, and 9%, respectively.

ABSTRAK

Pemprosesan Bahasa Semula jadi (NLP) adalah satu cara untuk komputer memperoleh, menganalisis, dan memahami makna bahasa manusia dengan cara yang pintar dan berguna. NLP menggunakan struktur hierarki bahasa yang membolehkan aplikasi dunia nyata seperti ringkasan teks automatik, penyelesaian peristiwa, pengekstrakan hubungan, dan pengiktirafan entiti untuk dipersembahkan dalam interaksi manusia-komputer yang tepat. Salah satu komponen NLP yang disebut sebagai Resolusi Rujukan Bersama (CR) adalah untuk menentukan sama ada dua frasa nama dalam bahasa semula jadi dapat merujuk kepada entiti yang sama. Dalam konteks ini, entiti boleh menjadi orang, organisasi, tempat, atau lain-lain, yang disebut sebagai istilah penyebutan entiti tersebut. Apabila tugas CR ini diperluaskan kepada beberapa dokumen, ia dipanggil sebagai Resolusi Rujukan Bersama Dokumen Silang (CDCR) yang memerlukan teknik khas untuk mengurus dan menangani rantai penyebutan dalam dokumen yang merujuk kepada entiti serupa di dokumen yang berbeza. Pada masa ini, terdapat beberapa limitasi dalam literatur yang ada di mana entiti CDCR yang dijana oleh varian rujukan tidak dapat dikenal pasti dengan baik, dan proses pengelompokan untuk membezakan entiti dengan kesamaan leksikal tidak ditangani dengan baik. Objektif utama penyelidikan ini adalah untuk mencadangkan satu model CDCR yang menggunakan penyisipan neural entiti dan penyebutan mereka, yang dijana dengan hanya menggunakan perkataan-perkataan daripada dokumen input. Model ini mencipta vektor penyebutan dan entiti menggunakan penyisipan neural, tanpa mengira penggunaan sumber luaran seperti Pangkalan Pengetahuan. Untuk pengelompokan entiti dan penyebutan yang lebih baik, teknik pengelompokan berdasarkan kepadatan yang diperbaiki yang mengandungi algoritma pengelompokan DBSCAN dan H-DBSCAN digunakan. Sebagai tambahan, satu prototaip bernama CROCER telah dirancang dan dibangunkan sebagai bukti konsep untuk menilai model dalam persekitaran eksperimen. Untuk penilaian, model ini diterapkan pada tiga set data yang tersedia untuk umum, yang disebut ‘John Smith Corpus’, ‘WePS-2 Collection’, dan ‘Google Wikilinks’ dari repositori sumber terbuka awam. Proses penilaian ini mengukur ketepatan, penarikan, dan skor F1 model oleh tiga sistem pemarkahan yang diketahui untuk Resolusi Rujukan Bersama iaitu MUC, B3, dan CEAF. Penemuan penyelidikan ini menunjukkan bahawa model yang dicadangkan dapat meningkatkan skor F1 dari set data masing-masing kepada 15.7%, 1.5%, dan 9%.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CDCR	-	Cross-Document Coreference Resolution
CR	-	Coreference Resolution
CWS	-	Contextual Words Sequence
ED	-	Entity Disambiguation
ER	-	Entity Resolution
ICR	-	Intra-document Coreference Resolution
IE	-	Information Extraction
IR	-	Information Retrieval
MWS	-	Mention Window Size
NER	-	Named Entity Recognition
NLP	-	Natural Language Processing
TWS	-	Training Window Size

LIST OF SYMBOLS

d	-	Document
D	-	Documents Set
e	-	Entity
E	-	Entity Set
m	-	Mention
M	-	Mention Group
t	-	Token

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

The mainstream part of the information produced by digital devices is globally expressed in the form of natural language text such as web pages, news articles, medical records, government documents, social media, etc. Such form of data is termed unstructured versus structured data. They are normalized and stored in a database that each record is divided from other records and relevant features that are associated with it. Information Extraction (IE) systems concern about automatically extraction of information from unstructured/semi-structured data (McCallum, 2005). For this purpose, to extract the locked information in unstructured text, Natural Language Processing (NLP) is used to discover and produce structured information.

NLP is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way. NLP considers the hierarchical structure of language: several words make a phrase, several phrases make a sentence and, ultimately, sentences convey ideas. By analyzing language for its meaning, NLP systems have long filled useful roles specially to analyze text, which allow machines to understand how human speak.

The field of NLP involves making computers to perform useful tasks with the natural language of human. NLP is characterized as a hard problem in computer science due to human language is rarely precise, or plainly spoken. To understand human language is to understand not only the words, but the concepts and how they are linked together to create meaning. Despite language being one of the easiest things for humans to learn, the ambiguity of language is what makes natural language processing a difficult problem for computers to master.

In NLP, there are various levels of ambiguity from Lexical Ambiguity which refers to the ambiguity of a single word to Pragmatic Ambiguity which refers to multiple interpretations of the text. To overwhelm the problems of NLP ambiguities, there are five general steps including Lexical Analysis, Syntactic Analysis, Semantic Analysis, Discourse Integration and Pragmatic Analysis.

Among various sub-tasks of NLP related to Discourse Integration (i.e., how the immediately preceding text's elements can affect the meaning and interpretation of the next elements). Coreference Resolution (CR) is essential to identify entity mentions in the text and resolve them into equivalent classes (H. Lee, Peirsman, Chang, Chambers, Surdeanu, & Jurafsky, 2011; Rahman & Ng, 2011b; Hajishirzi, Zilles, Weld, & Zettlemoyer, 2013; Màrquez, Recasens, & Sapena, 2013; Ng, 2016). In such context, an entity can be a real-world person, organization, or place, which is referred to, by a mention, i.e., a word or phrase referring to such an entity (Figure 1.1).

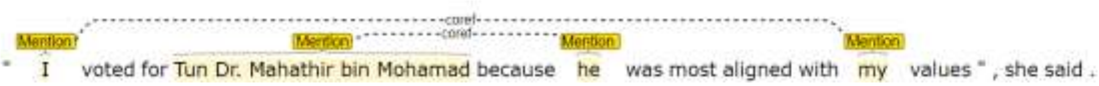

Figure 1.1 An Example of Coreference Resolution

The initial level of CR is to process the text within a single document, known as Intra-document Coreference Resolution (ICR) (Rao, McNamee, & Dredze, 2010). Expanding the scope of CR to process a collection of documents and resolving the entities across the documents leads to Cross-Document Coreference Resolution (CDCR) (Rao et al., 2010; Singh, Subramanya, Pereira, & McCallum, 2011; Ngomo, Röder, & Usbeck, 2014; Dutta & Weikum, 2015b; Beheshti, Benatallah, Venugopal, Ryu, Motahari-Nezhad, & Wang, 2016). CDCR plays a key role for several high-end NLP applications such as Automatic Knowledge Base Construction, Question Answering System, Automatic Text Summarization and Search Engines (Baron & Freedman, 2008; Dutta & Weikum, 2015b).

The reminder of this chapter consists of the different critical aspects of the research. Firstly, the background and statement of the problem are elaborated. This is followed by the research questions and objectives. Finally, the scope and significance of the research are briefly discussed.

1.2 Background of the Problem

CDCR consists of a variety of subtasks, starting with identifying mentions and entities and then co-referring them. The main goal is to group co-referring mentions to similar entities in clusters and distinct non-related mentions and entities. Mentions referring to the same entity are termed “coreferent” (Singh et al., 2011). CDCR can be viewed as a clustering problem of entity mention embedding based on their context similarities. However, local dependencies and entity contexts are ignored in standard clustering and high computational complexity is suffering as well. Accordingly, the main challenges of CDCR can be mentioned in three parts of, context detection dependencies, entity embedding for large datasets, and processing runtime of entity clustering.

For detecting the context of mentions and compute their similarities, pair-wise methods are used which are computationally expensive. Accordingly, such methods are unfeasible for CDCR tasks especially for large datasets. Furthermore, the entity disambiguation with similar strings or of the entity name variation should be enriched with precise detection of the mention contexts. While Knowledge Bases (KB’s) are employed in recent works (Hajishirzi et al., 2013; Dutta & Weikum, 2015b, 2015a) are used to enrich the relational information of entities, however, such featurization approaches cannot be reliable because the construction of KB’s depends on CDCR results.

Additionally**,** while machine learning algorithms are needed to be maintained with fixed-length inputs and produce fixed-length outputs, however text is not well-defined for such techniques. Word embedding is the collective name for a set of language modeling and feature learning techniques in NLP, where words or phrases from the vocabulary are mapped to vectors of real numbers (Goldberg, 2017). Considering the size of the feature vectors by recent techniques which depends on the vocabulary size of the document collection, by increasing in the number of mentions, the word embedding approaches based on them meet the problem with large datasets.

For the clustering step of CDCR task, it also meets two challenges. (1) Often the number of underlying entities and their identities are not known. (2) Unlike inference in other language processing tasks that scales linearly in the size of the corpus, the hypothesis dimension of features for coreference across documents grows super exponentially with the number of mentions. However, local dependencies and entity contexts are ignored in standard clustering and high computational complexity is suffering as well.

To handle the abovementioned problems, several solutions are proposed by researchers. Bagga and Baldwin (Bagga & Baldwin, 1998b) used the Vector Space Model (VSM) to disambiguate entities across documents. Later, Gooi and Allan (Gooi & Allan, 2004) presented three other models for CDCR based on the incremental vector space, KL divergence (the probabilistic approach), and a hierarchical clustering approach. More complicated models were presented by researchers later, established on one of the three main modelling approaches (Keshtkaran, Yuhaniz, & Ibrahim, 2017): graph-based model (Ngomo et al., 2014; Rahimian, Girdzijauskas, & Haridi, 2014; Emami, 2019), probabilistic model (Singh, Wick, & McCallum, 2010; Singh et al., 2011), and clustering-based model (Baron & Freedman, 2008; Finin, Syed, Mayfield, McNamee, & Piatko, 2009; Mayfield, Alexander, Dorr, Eisner, Elsayed, Finin, Fink, Freedman, Garera, & McNamee, 2009; Rao et al., 2010; Dutta & Weikum, 2015b). Using other approaches like streaming CDCR (Shrimpton, 2017), joint modeling of Cross-Document Entity and Event Coreference Resolution (Barhom, Shwartz, Eirew, Bugert, Reimers, & Dagan, 2019), and cross-lingual CDCR (Kundu, Sil, Florian, & Hamza, 2018) were also considered by researchers to use other external resources to outperform the results of CDCR. Nonetheless, they have not fully paved the way to satisfying results of resolving entities across documents regardless of any external information for any size of document collection.

Accordingly, while a few studies have been conducted in the area of CDCR, there are still open issues related to the CDCR task for processing effective context detection especially for large datasets. In order to address this goal, difficulties of large datasets for the number of records and dimension of the dataset, as well as effective context detection without conducting any contextual enrichment based on external sources should be considered. Therefore, the current research aims to design an improved model for CDCR task compared to the previous works which can outperform the effectiveness of the CDCR results.

1.3 Problem Statement

Identification and resolving co-referring entities across multiple documents by statistical data of the words and phrases of the document's text (i.e., frequency of the words or n-grams), provide useful data of the entity mentions and their context. Although such approaches deliver utilizable information of mention context to assist the differentiation of entities across documents, they are incapable of giving precise relationship between mentions and their context due to the ignorance of the sequence of words in the text. While, this problem is tried to be solved in recent works, however this procedure leads to a recursive dependency between CDCR task and KB's. This issue is produced due to the Automatic Knowledge Base Construction techniques which are relied on the results of CDCR. Accordingly, current techniques for CDCR task are suffering from limitations in independent context detection.

Other than the abovementioned issue, the CDCR task is facing with large datasets. The common approaches for embedding of mentions and their context (i.e., mapping words into numerical vectors) are heavily depended on the size of vocabulary of the data corpus. Increasing the size of vocabulary produces vectors with higher dimension in size and accordingly, will be more computationally expensive, time consuming or even impossible for the clustering analysis.

The problems of detecting the context of entities and their mentions in large datasets also produce the difficulties for the task of clustering of detected entities. Such problem becomes a critical issue together with the clustering challenge of CDCR (i.e., unknown number of clusters), which can raise the computation cost of the clustering task for CDCR.

1.4 Research Questions

Considering the aforementioned issues, this research aims to answer the following main question:

How to improve the effectiveness of detection and clustering of co-referent entities across multiple documents using only the documents' text, regardless of external information, for varied sizes of datasets?

In order to address the abovementioned question, three other questions are raised to answer that are defined in the following section. This research aims to answer the following questions:

(a) What are the existing approaches for detecting and clustering co-referent entities across documents?

(b) How to effectively construct the context of entities by merely document's text, regardless of external information?

(c) How to improve the effectiveness for the clustering task of detected entities by the proposed model?

(d) What is the improvement made by the proposed model for the results of CDCR?

1.5 Research Objectives

Based on the research questions, the research objectives are as follows:

(a) To identify the existing approaches for detecting and clustering co-referent entities across documents.

(b) To design a model for detecting the context of entities using the surrounding words of the mentions and their sequences regardless of external resources.

(c) To develop and improved clustering technique to enhance the effectiveness of CDCR.

(d) To evaluate the effectiveness of the model over benchmark datasets using standard metrics and comparing it results with the previous works.

1.6 Scope of the Study

The following research directions outline the boundaries of this study:

- **Coreference Resolution Across Documents:** While Coreference Resolution is about referring similar mentions in any kind of document set, the focus of this research is on Coreference Resolution across multiple documents as an advanced task against Coreference Resolution across the text of a document.

- **Document Types:** Source of text document can be any form like web pages, news articles, literary works, social media and so on. However, processing the text achieved from each source has its limitations. Generated text in social media may contain informal words, typo mistakes, or grammatical mistakes, literary text could be constructed with many literary terms, and web content may consist of many short phrases like titles, tables, or even in-complete sentences. Based on this, this research is limited to work on formal text which are almost certainly free of grammatical and typo mistakes and are made by complete sentences.

- **Entity Discovery:** This research focuses on Entity Discovery which is the task of clustering mentions into sets such that mentions in a given set all refer to the same real-world entity. Entity Discovery is against Entity Linking which is the problem of matching an entity with all of its referent mentions. The Entity Discovery is similar to Entity Linking, except it is more difficult because there are no known entities.

- **Entity Types:** Based on the definition by ACE (Automatic Content Extraction) which was a program of the early and mid of 2000's, entities are the most basic building blocks of the semantic representation. There are 7 types of entities: persons, organizations, GPEs (geo-political entities: locations with a government), [other] locations, facilities, vehicles, and weapons. Each entity has one or more mentions within the document. Each mention is either a name, a nominal, or a pronominal mention. However, this research is about resolving three main entity types, consist of Person, Organization and Location.

- **Cluster Analysis:** Coreferences within a document are generally based on rules or supervised learning using various kinds of linguistic features, such as syntactic paths between mentions, the distances between them, and their semantic compatibility as derived from co-occurrences. The CDCR task is essentially a clustering problem of entity embedding based on their context similarities. Based on this, this research is about learning the model for CDCR in an unsupervised manner regarding the contextual features of the text.

1.7 Assumptions and Limitations

Cross-Document Coreference Resolution (CDCR) is the task of identifying and co-referring similar entities across multiple documents. This task encompasses various kinds of activities and sub-tasks. Accordingly, the following assumptions and limitations are made in this research:

(a) CDCR consist of various stages which the initial is Intra-Document Coreference Resolution (ICR). In this research, this stage is conducted using a library called Stanford CoreNLP. This local CR stage may produce errors (e.g., incorrect chaining of mentions or omissions) which propagate the later stages. However, improving the result of ICR is out of the scope of this research.

(b) Based on the definition by ACE (Automatic Content Extraction) there are 7 types of entities: persons, organizations, GPEs (geo-political entities: locations with a government), [other] locations, facilities, vehicles, and weapons. However, this research is about resolving three main entity types, consist of Person, Organization and Location.

(c) The ICR sub-task may detect multiple entities from each document, related to the gold labels of the dataset or not. However, this research only concentrates on entities which their labels are provided in the dataset. Based on this, in the clustering stage, it is assumed that there is no outlier, and all of the entities will be included in one cluster.

(d) In the analytical phase for developing the model, it is assumed that the gold labels of benchmarking datasets are defined precisely. However, if any wrong or irrelevant gold label is found, it would be ignored.

(e) This research is only defined for applying the model on three selected datasets called, “John Smith Corpus”, “WePS-2 Collection”, and “Google Wikilinks” which are described in detail in Section 3.6.2.

1.8 Significance of the Study

By a new revolution in web search systems, user recommendations, and data analytics, transitioning from merely results of documents and keywords to knowledge and entities results is happening. Some instances of this phenomena are the IBM Watson technology, which is designed for deep question answering, and the Google Knowledge Graph and its applications. It seems that the most important value-adding part in this revolution is the identification and disambiguation of named entities in all of web and users’ contents.

These advances have been enabled by the creation of large knowledge bases (KB’s) such as DBpedia, Yago, or Freebase. Such semantic resources provide exceptionally large collections of entities like people, organizations, places, etc., which are enriched with more knowledge, describe their properties and relationships. In this situation, CDCR is a task which recognize and co-refer all mentions in an entire corpus that are related to the similar entity. CDCR does not involve mapping mentions to the entities of a KB, and unlike tasks like Named Entity Disambiguation, CDCR can deal with unknown or long-tail entities in KB's or even entities that are in very sparse form.

CDCR processes are also particularly important and have various applications in e-Health (processing the electronic health records), legal databases, opinions, sentiment analysis, and also understanding what is happening around us. Consider open-source intelligence as a motivating example, where millions of people broadcast events and opinions every second. In this context, cross document coreference occurs when the same person, place, event, or concept is discussed in more than one text source, e.g., tweets in Twitter. Consequently, CDCR can help in analyzing huge number of tweets generating in seconds, linking related tweets, and discovering more insight from them to understand what is happening now and predict what may happen later.

Designing and evaluating a suitable CDCR process are not only extremely important but also hugely challenging. Analyzing the state of the art, shows that a CDCR process involves multiple stages, where there are many possible choices for each stage, and only some combinations are valid.

1.9 Thesis Organization

This chapter fully discussed the nature of the research, the research gaps and problems faced, the research purpose and objectives, how these research gaps and problems will be addressed, as well as the research scope and significance. The remainder of this thesis is organized as in the second chapter a background on research directions, explains the unaddressed challenges, and presents a literature review of existing works on CDCR is described. The proposed research methodology is discussed in Chapter 3 by providing an overview of the research phases, operational framework, and explanations on benchmarking dataset and the validation and evaluation of these phases. The fourth chapter presents the research design and implementation by introducing the mathematical modeling of the CDCR process. The proposed techniques and algorithms are described in detail. The experimental results and a discussion are provided in Chapter 5 to indicate the applicability and feasibility of the proposed approach and investigate its evaluation and validation. Finally, a summary and conclusions of the thesis are provided in Chapter 6 by discussing the contributions of this research and suggesting for potential future research directions.

REFERENCES

Abeel, T., Peer, Y. V. d., & Saeys, Y. (2009). Java-ml: A machine learning library. *Journal of machine learning research, 10*(Apr), 931-934.

Ah-Pine, J., & Jacquet, G. (2009). *Clique-based clustering for improving named entity recognition systems.* Paper presented at the Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.

Allen, J. (1995). *Natural language understanding*: Pearson.

Ando, R. K., & Zhang, T. (2005). *A high-performance semi-supervised learning method for text chunking.* Paper presented at the Proceedings of the 43rd annual meeting on association for computational linguistics.

Ardanuy, M. C., van den Bos, M., & Sporleder, C. (2016). *You shall know people by the company they keep: person name disambiguation for social network construction.* Paper presented at the Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.

Artiles, J., Gonzalo, J., & Sekine, S. (2007). *The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task.* Paper presented at the Proceedings of the 4th International Workshop on Semantic Evaluations.

Artiles, J., Gonzalo, J., & Sekine, S. (2009). *Weps 2 evaluation campaign: overview of the web people search clustering task.* Paper presented at the 2nd web people search evaluation workshop (WePS 2009), 18th www conference.

Assal, H., Seng, J., Kurfess, F., Schwarz, E., & Pohl, K. (2011). *Semantically-enhanced information extraction.* Paper presented at the 2011 Aerospace Conference.

Attardi, G., Rossi, S. D., & Simi, M. (2010). *TANL-1: coreference resolution by parse analysis and similarity clustering.* Paper presented at the Proceedings of the 5th International Workshop on Semantic Evaluation.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The semantic web*, 722-735.

Bagga, A., & Baldwin, B. (1998a). *Algorithms for scoring coreference chains.* Paper presented at the The first international conference on language resources and evaluation workshop on linguistics coreference.

Bagga, A., & Baldwin, B. (1998b). *Entity-based cross-document coreferencing using the vector space model.* Paper presented at the Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1.

Baker, C. F. (2012). FrameNet, current collaborations and future goals. *Language resources and evaluation, 46*(2), 269-286.

Barhom, S., Shwartz, V., Eirew, A., Bugert, M., Reimers, N., & Dagan, I. (2019). Revisiting joint modeling of cross-document entity and event coreference resolution. *arXiv preprint arXiv:1906.01753*.

Baron, A., & Freedman, M. (2008). *Who is who and what is what: experiments in cross-document co-reference.* Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). paper recommender systems: a literature survey. *International Journal on Digital Libraries, 17*(4), 305-338.

Beheshti, S.-M.-R., Benatallah, B., Venugopal, S., Ryu, S. H., Motahari-Nezhad, H. R., & Wang, W. (2016). A systematic review and comparative analysis of cross-document coreference resolution methods and tools. *Computing, 99*(4), 313-349. doi:10.1007/s00607-016-0490-0

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research, 3*(Feb), 1137-1155.

Bengtson, E., & Roth, D. (2008). *Understanding the value of features for coreference resolution.* Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Black, W. J., Rinaldi, F., & Mowatt, D. (1998). *FACILE: Description of the NE System Used for MUC-7.* Paper presented at the Proceedings of the 7th Message Understanding Conference.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). *Freebase: a collaboratively created graph database for structuring human knowledge.* Paper presented at the Proceedings of the 2008 ACM SIGMOD international conference on Management of data.

Bollacker, K. D., Lawrence, S., & Giles, C. L. (1998). *CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications.* Paper presented at the Proceedings of the second international conference on Autonomous agents.

Boschee, E., Weischedel, R., & Zamanian, A. (2005). *Automatic information extraction.* Paper presented at the Proceedings of the International Conference on Intelligence Analysis.

Bryl, V., Giuliano, C., Serafini, L., & Tymoshenko, K. (2010). *Using Background Knowledge to Support Coreference Resolution.* Paper presented at the ECAI.

Campello, R. J., Moulavi, D., & Sander, J. (2013). *Density-based clustering based on hierarchical density estimates.* Paper presented at the Pacific-Asia conference on knowledge discovery and data mining.

Cao, Y. T., & Daumé III, H. (2019). Toward gender-inclusive coreference resolution. *arXiv preprint arXiv:1910.13913*.

Charikar, M., Chekuri, C., Feder, T., & Motwani, R. (1997). *Incremental clustering and dynamic information retrieval.* Paper presented at the Proceedings of the twenty-ninth annual ACM symposium on Theory of computing.

Chen, C., & Ng, V. (2012). *Combining the best of two worlds: A hybrid approach to multilingual coreference resolution.* Paper presented at the Joint Conference on EMNLP and CoNLL-Shared Task.

Chen, H.-H., Ding, Y.-W., & Tsai, S.-C. (1998). Named entity extraction for information retrieval. *Computer Processing of Oriental Languages, 12*(1), 75-85.

Chen, Y., & Martin, J. (2007). *Towards Robust Unsupervised Personal Name Disambiguation.* Paper presented at the EMNLP-CoNLL.

Clark, K., & Manning, C. D. (2016a). Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.

Clark, K., & Manning, C. D. (2016b). Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

Culotta, A., Kanani, P., Hall, R., Wick, M., & McCallum, A. (2007). *Author disambiguation using error-driven machine learning with a ranking loss function.* Paper presented at the Sixth International Workshop on Information Integration on the Web (IIWeb-07), Vancouver, Canada.

Culotta, A., Wick, M. L., & McCallum, A. (2007). *First-Order Probabilistic Models for Coreference Resolution.* Paper presented at the HLT-NAACL.

Das, A. S., Datar, M., Garg, A., & Rajaram, S. (2007). *Google news personalization: scalable online collaborative filtering.* Paper presented at the Proceedings of the 16th international conference on World Wide Web.

Daumé III, H., & Marcu, D. (2005). *A large-scale exploration of effective global features for a joint entity detection and tracking model.* Paper presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.

Davenport, J. H. (2018). Human-machine collaboration to disambiguate entities in unstructured text datasets. *Next-Generation Analyst VI, 10653*, 106530M.

Denis, P., & Baldridge, J. (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural, 42*(1), 87-96.

Domingos, P., Kok, S., Lowd, D., Poon, H., Richardson, M., & Singla, P. (2008). Markov logic. In *Probabilistic inductive logic programming* (pp. 92-117): Springer.

Domingos, P., & Lowd, D. (2009). Markov logic: An interface layer for artificial intelligence. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 3*(1), 1-155.

Dostert, L. E. (1955). The georgetown-ibm experiment. *1955). Machine translation of languages. John Wiley & Sons, New York*, 124-135.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). *Using latent semantic analysis to improve access to textual information.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Dutta, S., & Weikum, G. (2015a). *C3EL: A joint model for cross-document co-reference resolution and entity linking.* Paper presented at the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.

Dutta, S., & Weikum, G. (2015b). Cross-document co-reference resolution using sample-based clustering with knowledge enrichment. *Transactions of the Association for Computational Linguistics, 3*, 15-28.

Elsayed, T., Lin, J., & Oard, D. W. (2008). *Pairwise document similarity in large collections with MapReduce.* Paper presented at the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers.

Elsner, M., Charniak, E., & Johnson, M. (2009). *Structured generative models for unsupervised named-entity clustering.* Paper presented at the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Emami, H. (2019). A Graph-based Approach to Person Name Disambiguation in Web. *ACM Transactions on Management Information Systems (TMIS), 10*(2), 4.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.* Paper presented at the Kdd.

Estivill-Castro, V., & Houle, M. E. (2001). Robust distance-based clustering with applications to spatial data mining. *Algorithmica, 30*(2), 216-242.

Ferreira Cruz, A., Rocha, G., & Lopes Cardoso, H. (2020). Coreference resolution: toward end-to-end and cross-lingual systems. *Information, 11*(2), 74.

Finin, T., Syed, Z., Mayfield, J., McNamee, P., & Piatko, C. D. (2009). *Using Wikitology for Cross-Document Entity Coreference Resolution.* Paper presented at the AAAI Spring Symposium: Learning by Reading and Learning to Read.

Finkel, J. R., Grenager, T., & Manning, C. (2005). *Incorporating non-local information into information extraction systems by gibbs sampling.* Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.

Fleischman, M. B., & Hovy, E. (2004). *Multi-document person name resolution.* Paper presented at the Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics (ACL), Reference Resolution Workshop.

Garera, N., & Yarowsky, D. (2009). *Structural, transitive and latent models for biographic fact extraction.* Paper presented at the Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.

Ghoorchian, K., & Sahlgren, M. (2020). GDTM: Graph-based Dynamic Topic Models. *Progress in Artificial Intelligence, 9*, 195-207.

Giles, C. B., & Wren, J. D. (2008). Large-scale directional relationship extraction and resolution. *BMC bioinformatics, 9*(Suppl 9), S11.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies, 10*(1), 1-309.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Gooi, C. H., & Allan, J. (2004). *Cross-document coreference on a large scale corpus*. Retrieved from

Green, S., Andrews, N., Gormley, M. R., Dredze, M., & Manning, C. D. (2012). *Entity clustering across languages.* Paper presented at the Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Grishman, R., & Sundheim, B. (1996a). *Design of the MUC-6 evaluation.* Paper presented at the Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996.

Grishman, R., & Sundheim, B. (1996b). *Message Understanding Conference-6: A Brief History.* Paper presented at the COLING.

Haghighi, A., & Klein, D. (2007). *Unsupervised coreference resolution in a nonparametric bayesian model.* Paper presented at the Annual meeting-Association for Computational Linguistics.

Haghighi, A., & Klein, D. (2009). *Simple coreference resolution with rich syntactic and semantic features.* Paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3.

Hajishirzi, H., Zilles, L., Weld, D. S., & Zettlemoyer, L. S. (2013). *Joint Coreference Resolution and Named-Entity Linking with Multi-Pass Sieves.* Paper presented at the EMNLP.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

Harris, Z. S. (1964). Transformations in linguistic structure. *Proceedings of the American Philosophical Society, 108*(5), 418-422.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Holmes, D., & McCabe, M. C. (2002). *Improving precision and recall for soundex retrieval.* Paper presented at the Information Technology: Coding and Computing, 2002. Proceedings. International Conference on.

Hong, J. L. (2018). A Semantic-Based Document Checker. In *Redesigning Learning for Greater Social Impact* (pp. 205-213): Springer.

Huang, Y. (2019). *Automatic syntactic analysis of learner English.* University of Cambridge,

Hutchins, W. J. (1986). *Machine translation: past, present, future*: Ellis Horwood Chichester.

Jin, Y., Wu, D., & Guo, W. (2020). Attention-Based LSTM with Filter Mechanism for Entity Relation Classification. *Symmetry, 12*(10), 1729.

Joachims, T. (2006). *Training linear SVMs in linear time.* Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.

Jusoh, S. (2018). A study on NLP applications and ambiguity problems. *Journal of Theoretical & Applied Information Technology, 96*(6).

Kambhatla, N. (2004). *Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations.* Paper presented at the Proceedings of the ACL 2004 on Interactive poster and demonstration sessions.

Kang, N., Kim, J.-J., On, B.-W., & Lee, I. (2020). A node resistance-based probability model for resolving duplicate named entities. *Scientometrics, 124*(3), 1721-1743.

Kantor, B., & Globerson, A. (2019). *Coreference resolution with entity equalization.* Paper presented at the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

Karypis, G., & Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing, 20*(1), 359-392.

Keshtkaran, A., Yuhaniz, S. S., & Ibrahim, S. (2017). *An overview of cross-document coreference resolution.* Paper presented at the 2017 International Conference on Computer and Drone Applications (IConDA).

Kleenankandy, J., & KA, A. N. (2020). An enhanced Tree-LSTM architecture for sentence semantic modeling using typed dependencies. *Information Processing & Management, 57*(6), 102362.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*: MIT press.

Krishnamurthy, A., Balakrishnan, S., Xu, M., & Singh, A. (2012). Efficient active algorithms for hierarchical clustering. *arXiv preprint arXiv:1206.4672*.

Kundu, G., Sil, A., Florian, R., & Hamza, W. (2018). Neural cross-lingual coreference resolution and its application to entity linking. *arXiv preprint arXiv:1806.10201*.

Lavelli, A., Sebastiani, F., & Zanoli, R. (2004). *Distributional term representations: an experimental comparison.* Paper presented at the Proceedings of the thirteenth ACM international conference on Information and knowledge management.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics, 39*(4), 885-916.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). *Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task.* Paper presented at the Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task.

Lee, K., He, L., & Zettlemoyer, L. (2018). *Higher-Order Coreference Resolution with Coarse-to-Fine Inference.* Paper presented at the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).

Lin, X., & Chen, L. (2019). *Canonicalization of open knowledge bases with side information from the source text.* Paper presented at the 2019 IEEE 35th International Conference on Data Engineering (ICDE).

Loeliger, H.-A. (2004). An introduction to factor graphs. *IEEE Signal Processing Magazine*, *21*(1), 28-41.

Luo, X. (2005). *On coreference resolution performance metrics.* Paper presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., & Roukos, S. (2004). *A mention-synchronous coreference resolution algorithm based on the bell tree.* Paper presented at the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.

Ma, J., Liu, J., Li, Y., Hu, X., Pan, Y., Sun, S., & Lin, Q. (2020). *Jointly optimized neural coreference resolution with mutual attention.* Paper presented at the Proceedings of the 13th International Conference on Web Search and Data Mining.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9*(Nov), 2579-2605.

Maloney, C. J. (1962). Semantic information. *American documentation, 13*(3), 276-287.

Mann, G. S., & Yarowsky, D. (2003). *Unsupervised personal name disambiguation.* Paper presented at the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). *The Stanford CoreNLP natural language processing toolkit.* Paper presented at the Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). *The stanford corenlp natural language processing toolkit.* Paper presented at the ACL (System Demonstrations).

Màrquez, L., Recasens, M., & Sapena, E. (2013). Coreference resolution: An empirical study based on SemEval-2010 shared task 1. *Language resources and evaluation*, *47*(3), 661-694.

Mayfield, J., Alexander, D., Dorr, B. J., Eisner, J., Elsayed, T., Finin, T., . . . McNamee, P. (2009). *Cross-Document Coreference Resolution: A Key Technology for*

Learning by Reading. Paper presented at the AAAI Spring Symposium: Learning by Reading and Learning to Read.

McCallum, A. (2005). Information extraction: Distilling structured data from unstructured text. *Queue, 3*(9), 48-57.

McTear, M. F. (2016). *The rise of the conversational interface: A new kid on the block?* Paper presented at the International Workshop on Future and Emerging Trends in Language Technology.

Mendsaikhan, O., Hasegawa, H., Yamaguchi, Y., & Shimada, H. (2020). Quantifying the Significance and Relevance of Cyber-Security Text Through Textual Similarity and Cyber-Security Knowledge Graph. *IEEE Access, 8*, 177041-177052.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. A. (2019). Computing numeric representations of words in a high-dimensional space. In: Google Patents.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). *Distributed representations of words and phrases and their compositionality.* Paper presented at the Advances in neural information processing systems.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

Mohammad, S., & Hirst, G. (2006). *Determining Word Sense Dominance Using a Thesaurus.* Paper presented at the EACL.

Mousavi, H., Kerr, D., Iseli, M., & Zaniolo, C. (2014). *Mining semantic structures from syntactic structures in free text documents.* Paper presented at the Semantic Computing (ICSC), 2014 IEEE International Conference on.

Murray, T. E. (1995). *The structure of English: Phonetics, phonology, morphology*: Allyn and Bacon.

Nastase, V., Strube, M., Boerschinger, B., Zirn, C., & Elghafari, A. (2010). *WikiNet: A Very Large Scale Multi-Lingual Concept Network.* Paper presented at the LREC.

Ng, V. (2008). *Unsupervised models for coreference resolution.* Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Ng, V. (2010). *Supervised noun phrase coreference research: The first fifteen years.* Paper presented at the Proceedings of the 48th annual meeting of the association for computational linguistics.

Ng, V. (2016). Advanced Machine Learning Models for Coreference Resolution. In *Anaphora Resolution* (pp. 283-313): Springer.

Ng, V., & Cardie, C. (2002). *Improving machine learning approaches to coreference resolution.* Paper presented at the Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.

Ngomo, A.-C. N., Röder, M., & Usbeck, R. (2014). *Cross-document coreference resolution using latent features.* Paper presented at the Proceedings of the Second International Conference on Linked Data for Information Extraction-Volume 1267.

Ni, Y., Zhang, L., Qiu, Z., & Wang, C. (2010). Enhancing the open-domain classification of named entity using linked open data. In *The Semantic Web–ISWC 2010* (pp. 566-581): Springer.

Niu, C., Li, W., & Srihari, R. K. (2004). *Weakly supervised learning for cross-document person name disambiguation supported by information extraction.* Paper presented at the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.

Oshika, B. T., Evans, B., Machi, F., & Tom, J. (1988). *Computational techniques for improved name search.* Paper presented at the Proceedings of the second conference on Applied natural language processing.

Pantel, P., Crestan, E., Borkovsky, A., Popescu, A.-M., & Vyas, V. (2009). *Web-scale distributional similarity and entity set expansion.* Paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2.

Papagiannopoulou, E., & Tsoumakas, G. (2020). A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(2), e1339.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*: Elsevier.

Raff, E. (2017). JSAT: Java statistical analysis tool, a library for machine learning. *The Journal of Machine Learning Research, 18*(1), 792-796.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., & Manning, C. (2010). *A multi-pass sieve for coreference resolution.* Paper presented at the Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

Rahimian, F., Girdzijauskas, S., & Haridi, S. (2014). *Parallel community detection for cross-document coreference.* Paper presented at the Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on.

Rahman, A., & Ng, V. (2009). *Supervised models for coreference resolution.* Paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2.

Rahman, A., & Ng, V. (2011a). *Coreference resolution with world knowledge.* Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.

Rahman, A., & Ng, V. (2011b). *Ensemble-based coreference resolution.* Paper presented at the IJCAI Proceedings-International Joint Conference on Artificial Intelligence.

Ramshaw, L., Boschee, E., Bratus, S., Miller, S., Stone, R., Weischedel, R., & Zamanian, A. (2001). *Experiments in multi-modal automatic content extraction.* Paper presented at the Proceedings of the first international conference on Human language technology research.

Rao, D., McNamee, P., & Dredze, M. (2010). *Streaming cross document entity coreference resolution.* Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters.

Ratinov, L., & Roth, D. (2012). *Learning-based multi-sieve co-reference resolution with knowledge.* Paper presented at the Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Ravichandran, D., Pantel, P., & Hovy, E. (2005). *Randomized algorithms and nlp: using locality sensitive hash function for high speed noun clustering.* Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.

Recasens, M., de Marneffe, M.-C., & Potts, C. (2013). *The life and death of discourse entities: Identifying singleton mentions.* Paper presented at the Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*: Cambridge university press.

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine learning, 62*(1-2), 107-136.

Riedl, M., Betz, D., & Padó, S. (2019). *Clustering-based article identification in historical newspapers.* Paper presented at the Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science, 290*(5500), 2323-2326.

Sarmento, L., Kehlenbeck, A., Oliveira, E., & Ungar, L. (2009). An approach to web-scale named-entity disambiguation. In *Machine Learning and Data Mining in Pattern Recognition* (pp. 689-703): Springer.

Shrimpton, L. W. (2017). Efficient techniques for streaming cross document coreference resolution.

Singh, S., Subramanya, A., Pereira, F., & McCallum, A. (2011). *Large-scale cross-document coreference using distributed inference and hierarchical models.* Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.

Singh, S., Subramanya, A., Pereira, F., & McCallum, A. (2012). Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2012-015*.

Singh, S., Wick, M., & McCallum, A. (2010). Distantly labeling data for large scale cross-document coreference. *arXiv preprint arXiv:1005.4298*.

Singhal, A. (2012). Introducing the knowledge graph: things, not strings. *Official google blog, 5*.

Sivic, J., & Zisserman, A. (2008). Efficient visual search of videos cast as text retrieval. *IEEE transactions on pattern analysis and machine intelligence, 31*(4), 591-606.

Skut, W., & Brants, T. (1998). Chunk tagger-statistical recognition of noun phrases. *arXiv preprint cmp-lg/9807007*.

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics, 27*(4), 521-544.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). *Yago: a core of semantic knowledge.* Paper presented at the Proceedings of the 16th international conference on World Wide Web.

Sundheim, B. M. (1996). *Overview of results of the MUC-6 evaluation.* Paper presented at the Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996.

Syed, Z. S., Finin, T., & Joshi, A. (2008). *Wikitology: Using wikipedia as an ontology.* Paper presented at the proceeding of the second international conference on Weblogs and Social Media.

Tabebordbar, A. (2020). Augmented Understanding and Automated Adaptation of Curation Rules. *arXiv e-prints*, arXiv: 2007.08710.

Tasdemir, K., & Merényi, E. (2011). A validity index for prototype-based clustering of data sets with complex cluster structures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41*(4), 1039-1053.

Tauer, G., Date, K., Nagi, R., & Sudit, M. (2019). An incremental graph-partitioning algorithm for entity resolution. *Information Fusion, 46*, 171-183.

Team, D. (2016). Deeplearning4j: Open-source distributed deep learning for the jvm. *Apache Software Foundation License, 2*.

Toruk, M., Bilgin, G., & Serbes, A. (2020). *Speaker Diarization using Embedding Vectors.* Paper presented at the 2020 28th Signal Processing and Communications Applications Conference (SIU).

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). *Feature-rich part-of-speech tagging with a cyclic dependency network.* Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the

Association for Computational Linguistics on Human Language Technology-Volume 1.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., & Tsujii, J. i. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Advances in informatics* (pp. 382-392): Springer.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). *A model-theoretic coreference scoring scheme.* Paper presented at the Proceedings of the 6th conference on Message understanding.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing, 17*(4), 395-416.

Vrandečić, D. (2012). *Wikidata: A new platform for collaborative data collection.* Paper presented at the Proceedings of the 21st international conference on world wide web.

Wauthier, F. L., Jojic, N., & Jordan, M. I. (2012). *Active spectral clustering via iterative uncertainty reduction.* Paper presented at the Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.

Wellner, B., McCallum, A., Peng, F., & Hay, M. (2004). *An integrated, conditional model of information extraction and coreference with application to citation matching.* Paper presented at the Proceedings of the 20th conference on Uncertainty in artificial intelligence.

Wen, S., Wei, H., Yang, Y., Guo, Z., Zeng, Z., Huang, T., & Chen, Y. (2019). Memristive LSTM network for sentiment analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Wick, M., Rohanimanesh, K., Culotta, A., & McCallum, A. (2009). Samplerank: Learning preferences from atomic gradients. *Advances in Ranking*, 69.

Wick, M., Singh, S., & McCallum, A. (2012). *A discriminative hierarchical model for fast coreference at large scale.* Paper presented at the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1.

Wick, M. L., Culotta, A., Rohanimanesh, K., & McCallum, A. (2009). *An Entity Based Model for Coreference Resolution.* Paper presented at the SDM.

Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. In: Sage Publications Sage UK: London, England.

Yang, S., & Wei, R. (2020). Semantic Interoperability Through a Novel Cross-Context Tabular Document Representation Approach for Smart Cities. *IEEE Access, 8*, 70676-70692.

Yang, X., Su, J., Zhou, G., & Tan, C. L. (2004). *Improving pronoun resolution by incorporating coreferential information of candidates.* Paper presented at the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.

Zawisławska, M., Ogrodniczuk, M., & Szczyszek, M. (2021). Indirect Relations and Frames: Coreference in Context.

Zhang, R., dos Santos, C., Yasunaga, M., Xiang, B., & Radev, D. (2018). *Neural Coreference Resolution with Deep Biaffine Attention by Joint Mention Detection and Mention Clustering.* Paper presented at the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

Zhang, T., Li, H., Ji, H., & Chang, S.-F. (2015). *Cross-document Event Coreference Resolution based on Cross-media Features.* Paper presented at the EMNLP.

Zheng, J., Vilnis, L., Singh, S., Choi, J. D., & McCallum, A. (2013). Dynamic knowledge-base alignment for coreference resolution.

Appendix A Alphabetical List of Part-of-speech Tags Used in The Penn Treebank Project

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund, or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

LIST OF PUBLICATIONS

Non-Indexed Journal

1. **Aliakbar, K.**, Siti Sophiayati, Y., & Mohammad Reza, R. (2019). Distributed Representation of Entity Mentions Within and Across Multiple Text Documents. Open International Journal of Informatics (OIJI), 7(1), 35-46. http://apps.razak.utm.my/ojs/index.php/oiji/article/view/195.

Non-Index Conference Proceedings

1. **Aliakbar, K.**, Siti Sophiayati, Y., & Suhaimi, I. (2017). An overview of cross-document coreference resolution. In *2017 International Conference on Computer and Drone Applications (IConDA)* (pp. 43-48). IEEE. https://doi.org/10.1109/ICONDA.2017.8270397.