Privacy in Text Documents

[1,2]Mariana DIAS, [1,2,3] João C. FERREIRA, [2]Rui MAIA,[2]Pedro SANTOS, [1]Ricardo RIBEIRO
[1]Instituto Universitário de Lisboa (ISCTE-IUL), [2]Inov Inesc Inovação - Instituto De Novas Tecnologias, [3] Information Sciences, Technologies and Architecture Research Center (ISTAR-IUL)
mrdso@iscte-iul.pt

Abstract: The process of sensitive data preservation is a manual and a semi-automatic procedure. Sensitive Data preservation, in particular, the handling of confidential, sensitive and personal information, suffers from many problems. The sensitive data identification in documents still requires human intervention which can be a very tedious and prone to errors process. DataSense will be a highly exportable software that will enable organizations to identify and understand the sensitive data in their possession in unstructured textual information (digital documents) to comply with legal, compliance and security purposes. The goal is to identify and classify sensitive data (Personal Data) present in large-scale structured and non-structured information in a way that allows entities and/or organizations to understand it without taking into question security or confidentiality issues, allowing companies that focus on their clients to better understand their profile from information collected from sensitive data. The DataSense project will be based on European-Portuguese text documents with different approaches of NLP (Natural Language Processing) technologies and the advances in machine learning, such as Named Entity Recognition and Disambiguation. It will also be characterized by the ability to assist organizations in complying with standards such as the GDPR (General Data Protection Regulation), which regulate data protection in the European Union.

Keywords: Sensitive Data, Natural Language Processing, Text Mining, Named Entities Recognition.

I. Introduction

In the context of an information society where more and more documents are generated and collected from various sources and by various entities, it is only natural that this situation raises more and more security concerns. The complexity and severity of security issues in systems and even related to individuals depend crucially on how organizations deal with sensitive data of any kind. These are problems that have worsened over time in a fully digitized society which generates large-scale amounts of information that easily leads to a loss of control over the content of these documents. In the past, as an example for documents that needed to be public, the data considered to be sensitive to an entity or individual and abstracted by manual procedures, duly documented and structured using fixed rules in a process called "sanitization", were manually identified. More recently, tools have been created that help the identification process with a particular focus on structured information such as emails, addresses, phone numbers or credit cards, while leaving all sensitive data of a textual and unstructured nature as is the case of names, medical information, criminal records, religion to the care of human expertise to identify and treat them. All this manual and semi-automatic process suffer from several problems, namely:
Identifying sensitive data in one document (or several) requires tasks that are manual, error-prone, and therefore very costly;
Their identification in large-scale documents (e.g. thousands of documents) does not allow an approach that depends on human expertise in their identification and relationship in most cases;
Since the "identification" of sensitive data makes up an important part of the whole process (even if one uses only human expertise), this is only part of an even bigger problem. Incorrect management of this type of data can put public or private organizations in very complex situations even in illegal situations. To fight such situations, it is necessary for organizations to have the means to detect them and to carry out, in parallel, integrated management of all sensitive data following existing standards and legislation. For this, it is essential that organizations and entities can perceive "where" they are, what "type" they are and "how" they relate the data they have. Only with a strong understanding of sensitive data, namely their identification, classification and the relationships they have (regardless of their format) will it allow organizations to deal with a problem that has become too complex and expensive. This understanding, once obtained, allows organizations to perceive, create and systematize preventive security policies, educate users in their manipulation, set tight controls for sensitive data and implement rules following current legislation. There are mandatory responses that will have to be given by entities and organizations to a number of existing issues, such as the right to forgetfulness, the request for access to personal data stored by users, temporary authorizations to store and process personal and sensitive data, as well as, the automatic processing of sensitive information.
In our work, we try to create a platform that allows acting in the area of data discovery that is considered Sensitive (Sensitive Data Discovery). DataSense about data privacy has two fundamental objectives:
1. Allow the identification and classification of sensitive data present in unstructured information on a large scale in order to allow entities and organizations to obtain an understanding of their sensitive data;
2. Allow organizations to respond immediately to the content and network (direct and indirect relationships) of the sensitive data they store and process (e.g., right to forget).
In order to respond to the aforementioned objectives, our platform is based on Named Entity Recognition and Classification essential to overcome the state of the art of application and proposes a hybrid architecture that will take the risk of applying the area of Natural Language Processing (NLP) and Automatic Learning (Machine Learning) in the critical area of sensitive data protection.

II. Related Concepts and Work

In the banalization of the commercial discourse on AI solutions, there was a considerable growth of business investment in the most diverse sub-areas of this topic, which does not escape NLP and where this platform is located. NLP is used in numerous business applications ranging from personal assistants in smartphones to real-time translation systems and social-emotional analysis. More recently, NLP has begun to be expanded to incorporate more mature models with better levels of efficiency and precision, and the result is more intelligent and capable applications. In the commercial area, the use of NLP is at an advanced level for identification and classification of sensitive data, but the application in the Portuguese language is not an area that has developed sufficiently.

The concept of Sensitive Data or Sensitive Information can follow several points of view depending on the context and on purpose, is often linked to tasks of data anonymization. Most systems that perform anonymization of sensitive data work in four steps: (i) pre-processing, (ii) detection of sensitive data, (iii) classification of sensitive data and (iv) anonymization.
Thus, different proposals appear in order to deal with the automatic discovery of sensitive data and the extraction of information. There are several examples of applications that work in the area of eDiscovery (Electronic Discovery). Cicayda, for example, looks for documentary information, legal information data to catalogue and perform a risk analysis using non-detailed natural language techniques. Another solution in the market is called Onna that allows a search in different repositories but uses standards techniques only to detect unstructured information found and catalogued. Both solutions and essentially generic systems of Electronic Discovery that classify only the metadata of the documents and not their content and in some cases with straightforward approaches to NLP such as regular expression processing. Another problem associated with these systems is that they do not allow its use in other languages like Portuguese for instance. In this context and knowing that Portuguese is a language with more than 200 million speakers in the world, it is essential to consider it for these type of systems.
However, there are much more advanced possibilities that can be applied and that our platform integrates into its approach. Some of the most advanced concepts in this technical-scientific area and some of the open challenges are described below.
Progress in the area of AI, particularly in the area of NLP technologies has been notable, with visible effects on the quantity and quality of products, systems and applications based on natural interaction. Firstly, because there are areas where the sensitivity of information is decisive and any error can have serious consequences. For example, it is not possible to apply massively and easily natural language processing systems in the legal area or the medical field. To solve this type of problems implies necessarily the ability to extract information, classified information and identify documents in large databases and relational documentary information.

There is an insufficient number of the corpus or annotated data sets to train and validate this type of systems in the European Portuguese language and the specific area of sensitive information. Resources in European Portuguese are generally much more limited than those in languages such as English, and therefore there are not many production systems based on natural language processing in PT-EU, but there are some studies by Fonseca et al. (2014) and de Souza et al (2008). The context and challenges of the applications supported by AI, namely in the area of NLP are evidenced in the solutions of Information Extraction and Retrieval and Named Entity Recognition. These are aimed at obtaining the semantic structure – the objects, their relations and actions from data in written natural language, which in the most complex cases may not be structured.
In addition to the challenges inherent in the complexity of large documentary systems with various data sources in various formats, under typically poor quality, morphosyntactic variability is added. The different ways of writing a sentence and the ambiguity of the natural language itself (different meanings of a word, expression or phrase) characterize a high level of complexity in the development of a solution based on NLP.

In the context of this platform, it is important to mention that it is very relevant to identify and define the fundamental ontology. This should be distinguished from ontologies used in other domains and languages such authors in Weischedel et al (2013) by the integration of semantic knowledge in the area of sensitive data in the area of information structuring, namely extraction and retrieval.
Natural Language is a common component of all AI applications based on natural language understanding. Most NLP-based projects for specific domains use rule-based modules as authors in Ronan and Weston (2008), such as regular expressions and syntax rules. However, this approach entails two problems: 1) the system only recognizes a limited set of rules; 2) extension or improvement requires manual labour of someone who knows the domain and the formalism of rule-making. Other techniques, based on comparison with lexicons or word dictionaries, or also using Machine Learning approaches in the research study by Lample et al, can learn to classify or even generate new rules but assume the existence of a known annotated corpus.
In general, in Named Entity Recognition not only entities are identified, but also classified according to a given set of types. For example, the well-known shared task of CoNLL Sang and Meulder (2003) divided groups of named entities into three classes (organizations, places and people).
Named Entity Recognition consists in identifying terms in a specific text composed by one or more tokens as mentioned in research Nadeau and Sekine (2007). The most common types of named entities are Names (Personal Names, Names of Organizations ...), Locals, or Personal Information (mobile phone number, identification number, postal code...)
The recognition task also implies the classification of the data type, within the various categories. This technique consists of identifying keywords in the text of documents.

The HAREM conference, Santos and Cardoso (2008), is an evaluation event for Portuguese Language that aims to develop systems of Named Entities Recognition. In addition to the rules-based approaches as done in Collobert and Weston (2008) and Wiseman et al (2015), there are many used automatic statistical approaches with Conditional Random Fields (CRF) models as was mentioned in John Lafferty et al (2001), or as Hidden Markov Models (HMM) in Ponomareva et al (2007), Maximum Entropy Markov Models (MEMM) studied by Borthwick et al (1998).

Subsequently, beyond statistical approaches, neural networks presented in Palangi et al (2016) studies show that these can be trained for different types of data and domains.
It has had successive developments being the target of recent application of approaches based on latent structures in research by Martschat and Strube (2015), or reinforcement learning by Clark and Manning (2016), for example. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) neural networks, and taking into account the specific linguistic knowledge of a given language and domain as Dhingra et al (2017) mentioned, are considered as improving the results this complex area.
The best results are achieved through hybrid approaches by combining rules-based methods and Machine Learning techniques.

III. Proposal

The Platform solution defines and integrates three fundamental concepts that are integrated into the field of extraction and retrieval of sensitive data present in large unstructured databases. The concepts and the hybrid approach are detailed below.

Concept 1 - Sensitive Data (Personal Data).
Despite the convenience of using the acronym PII (Personal Identification Information), in the European context there is something that is not a direct synonym, but something called PD (Personal Data). Personal Data is supported by three different Directives: 95/46 / EC – Data Protection Directive that was replaced in May 2018 by the General Data Protection Regulation (GDPR) Directive; 2002/58 / EC (E-Privacy Directive which was also replaced in May 2018 by E-Privacy Regulation); and, 2006/24 / EC - Article 5 (Data Retention Directive). The Data Protection Directive such as the GDPR regulates the processing of personal data in the European Union. The GDPR directive, which applies a set of rules to return control of sensitive data to citizens and also establish clear and objective rules on deadlines, mechanisms and penalties for non-implementation. The E-Privacy directive, which was replaced at the same time as the Data Protection Directive by E-Privacy Regulation, defines rules on confidentiality of information, treatment of spam, handling of cookies, etc. Finally, Directive 2006/24 / EC regulates data retention, in particular in the telecommunications industry.

For this work, will be defined as a structure that represents the sensitive data that can identify, contact, or locate an individual. This crucial work will be carried out based on the best practices and accumulated know-how in the area of Sensitive Data. The basic structure of information can be described already, and generically, through the following information contexts: 1) Identification personal information (name, email, social security number, credit card number, etc.); 2) Information about locals (residence, workplace…); 3) Information about entities, organizations or companies directly associated (e.g. employment) with the individual;

Concept 2 - Natural Language Processing.
NLP is a subarea of AI that studies the ability and limitations of a machine to understand and generate natural, spoken or written, language. The purpose of Natural Language Understanding is to provide computers with the ability to understand texts that humans easily understand and interpret but often do not follow the formal, syntactic and semantic characteristics of grammar and definitions considered to be formally correct in that language. "Understanding" a text means recognizing the context, performing lexical and morphological, syntactic, and semantic, analysis, creating abstracts, extracting information, interpreting the senses, analyzing feelings, and even learning concepts with processed texts. Thus, the use of NLP, and specifically Natural Language Understanding, is mandatory in the solution and innovative in this area of application – namely for European Portuguese – given that it will allow, through trained NLP models, to perform the identification and classification of sensitive data. For this will be applied a set of NLP tools (tokenization, part-of-speech tagging, lemmatization, stemming...)

Concept 3 - Multi-format and unstructured information analysis.
The main goal is to build a solution capable of identifying and classify text present in documents, emails and all types of databases or information repositories. To achieve this mechanism, it will be applied several techniques of the document and text analysis having the goal of extracting all the sensitive data found along these information sources.

IV. Technical-Scientific Approach to the Solution

The DataSense solution intends to aggregate previously defined concepts (1 to 3) into a simple and scalable system that is capable of fulfilling all the objectives it proposes. The tool will have the capacity to process large amounts of data and will achieve a level of accuracy close to the human being. For this, we describe here the essential steps in the design and implementation of the system based on NLP for Information Extraction and Classification.
Using a hybrid approached we identify Named Entities based in Regular Expressions, using lexicons and dictionaries and Machine Learning techniques to identify the following types of data:

- Civil Identification Number
- Bank Identification Number
- Credit Card Number
- Tax Identification Number
- Passport Number
- Social Security Number
- Health User Number
- Telephone Contact
- Driving License Number
- email
- Postal code
- Person Names
- Addresses
- Locals
- Organizations/Institutions

a) Data

The HAREM named entity data Ferrández et al (2007), consists of a set of documents covering the Portuguese Language. HAREM data has a golden collection consisting of 129 text documents of different genres such as: News, Interviews, Blogs, Publicity Texts, Web Pages, etc. This data set is annotated for ten different categories distributed in an unbalanced way as we can see in table 1 that illustrates the number of occurrences for each category.

Table 1. Number of Occurrences per Category

Category	Number of Occurrences
PERSON	2 036
LOCAL	1 311
TIME	1 180
ORGANIZATION	961
WORK	449
VALUE	353
COISA	308
EVENT	300
ABSTRACTION	286
OTHER	79
total	7 263

For DataSense only the following categories were considered: LOCAL, ORGANIZATION and PERSON, reducing the classification to only four groups.
The annotation of the HAREM dataset is made according to the XML format, contains tags and additional information that was not used in this work. All annotations start with the “EM” tag and end with “/ EM”, as well as an ID attribute for easy identification. An example of an annotation is:
<EM ID="a55968-47" CATEG="PESSOA" TIPO="CARGO"> Presidente da Câmara de Nova Iorque</EM>

Since the data in XML format does not serve as input, in order to feed the algorithms the dataset had to be transformed considering only the CATEG tag. It was necessary to transform the data into CoNLL, Sang and Meulder (2003) format with IOB tags which means:

- I (Inside) means that the current token belongs to the entity;
- O (Out) means that the current token does not belong to the entity;
- B (Begin) means that the current token is the first of the entity.

In addition to the HAREM dataset, we have used dictionaries of Person Names, Locals and Organizations.

b) Data preprocessing

As a way to improve the obtained results in its applied data pre-processing techniques for the document's texts. This pre-processing consists only in basic text processing tasks, tokenization (i.e., separation of phrases in n-grams) such as in research Teixeira et al (2011) and text segmentation, separating punctuation marks into individual items, respecting the language rules, part-of-speech tagging to the entire text, associated each n-gram a tag with additional information.

c) Evaluation

In the context of the project, we also evaluated some metrics of the methods and models (precision, recall, F1- measure) that allows assessing the performance in the identification and classification of sensitive data. Precision is the percentage of named entities found by the learning system that is correct. The recall is the percentage of named entities present in the corpus that are found by the system. F1- measure is a measure of a test's accuracy, it considers both the precision and the recall. A named entity is correct only if it is an exact match of the corresponding entity in the data file. The decision of which metrics to use will check for those with higher accuracy, fewer false positives, less ambiguity and metonymy.

V. Implementation and Results

After analyzing the data, HAREM dataset and the word dictionaries found for the Portuguese language. In order to identify and classify Named Entities, DataSense implements a hybrid language processing approach for the Portuguese idiom.
For some categories, we decided to use more than one approach and choose the one with the better results to integrate into DataSense platform. The implementation combined different NLP tasks and was divided into three distinct parts:

a) Ruled Base

Ruled Base approach to recognize personal identification numbers (Civil Identification, Bank Identification, Credit Card, Tax Identification, Passport, Social Security, Health User, Telephone Contact, Driving License), the chosen approach based in regular expressions.
These types of sensitive data can easily be identified solely through surface structure patterns. Rule-based NER systems can be very effective, but require some manual effort and should only be used in very specific areas like this.

b) Application of the Dictionaries

For the Persons Name category, the application of the dictionaries was one of the used approaches. We use two types of dictionaries one with all female names registered in the last five years and another with male names. Using additional contexts such as capital letters and POS and editing measures, such as the minimum editing distance, Ristad and Yianilos (1998) that will allow you to check for spelling mistakes or alternative ways of writing.
In this approach, every word, with POS tag "NOUN" and capital letter, from each phrase was compared.

c) *Machine* Learning

In this experiment, two implementations were taking into place, to compare results and test with different methodologies. In these experiences, the objective was the Recognition Named Entities of the categories Names, Locals and Organizations.
The first implementation consists of a CRF (Conditional Random Fields) probabilistic model, the approach taken by Lample et al for the Recognition of Enumerated Entities. The second experience with a statistical approach with the Spacy library for the Entity Recognition mentioned, following the approach taken by Pires (2017).

The results for the different approaches for the different categories (PERSON, LOCAL and ORGANIZATION) can be seen in table 2, 3 and 4. Tables show the obtained scores by the system through different stages of the analysis. All tests were performed using the HAREM dataset annotated for those categories.
We evaluated the same three approaches — dictionaries application, CRF models and Spacy library —, using 70% train and 30% test. The tested hyperparameters were the default ones, and the best ones found in the hyperparameter study.

In table 2 can see the results for the Person Names category using the dictionaries application approach, referred in section b).

Table 2. Person category results for dictionaries application

	Precision	Recall	F1-measure
PERSON	0.59	0.31	0.41

Table 3 shows the results of the CRF model approach. Comparing with the previous result (table 2) we can see that the Person category had better results. The results compared to the three categories are quite similar.

Table 3. A person, Local and Organization categories result using CRF Model

	Precision	Recall	F1-measure
PERSON	0.67	0.86	0.76
LOCAL	0.67	0.65	0.66
ORGANIZATION	0.66	0.58	0.62

Table 4 shows the results for the Spacy library application. With the analysis of the results, we can see that the results obtained through the CRF model were better in all categories. This is because in this experience the use of Spacy library was taking care without having any context or training data to the model.

Table 4. A person, Local and Organization categories result using Spacy library

	Precision	Recall	F1-measure
PERSON	0.63	0.59	0.61
LOCAL	0.41	0.53	0.47
ORGANIZATION	0.40	0.67	0.50

When analyzing the results of the previous tables, we can see that the best results were achieved with the training of a CRF model.The Precision results are on average 12% higher than those obtained with the other approaches. This although, does not mean that CRF model is best for this type of task. These results were obtained with the Spacy library through an experiment without any kind of context, making only use of itself. In the case of the results of table 2, this would certainly be better if the set of present names in the applied dictionaries were larger.

VI. Conclusion

We describe our work towards a severe problem of sensitive information. As a way to give answers that allow to understand and define controls for the sensitive data and to solve the automatic processing of sensitive information, the proposed system was implemented using NLP and Machine Learning techniques.

DataSense is a functional prototype capable of correctly executing a set of tasks that must obey a set of objectives. Objectives can be divided into two components of the system: Recognition and Classification of Named Entities.

With this work, we achieved to have a system that is capable to correctly identify the proposed sensitive data achieving a significant result for the Portuguese idiom. The developed work is capable of being used in several corporate domains to protect GDPR and be replicated to other types of data and different categories. The obtained results could have been better through the use of bigger training datasets, which would allow training other types of models, such as LSTM neural networks. Adding more context to the identification of named entities, or through a morphological analysis and more detailed disambiguation could produce an increase of precision, and this will be one of the next stages of our work.

Regarding the evaluation metrics and the results that we obtained when compared with the systems that used the same dataset (HAREM), we obtained results quite close with current state of the art. Being that our work stands out by the implementation of rules for the recognition of all the sensitive information referred in IV chapter. It was not possible to make an evaluation by comparison to this type of data, since there are no available datasets but through an exhaustive and detail analysis we have managed to conclude an excellent operation of our system at the level of these categories of sensitive data

Acknowledgement

The DataSense project has is being led by Link Consulting and received co-funding from the FEDER - Lisbon 2020, PT 2020, European Union’s PT 2020 research and innovation program under grant agreement cod POCI-01-0247-FEDER-038539.

References

Borthwick, A., Sterling, J., Agichtein, E. and Grishman, R. (1998). "Exploiting diverse knowledge sources via maximum entropy in named entity recognition". In Proceedings ofthe Sixth Workshop on Very Large Corpora, pages 152–160

Clark, K. and Manning, C. D. (2016). “Deep reinforcement learning for mention-ranking coreference models,” ArXiv Prepr.

Collobert, R. and Weston, J. (2008) “A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning,” in Proceedings of the 25th International Conference on Machine Learning, New York, NY, USA, pp. 160–167.

de Souza, J. G. C., Gonçalves, P. N. and Vieira, R. (2008, September). "Learning coreference resolution for portuguese texts". In International Conference on Computational Processing of the Portuguese Language (pp. 153-162). Springer, Berlin, Heidelberg.

Dhingra, B., Yang, Z., Cohen, W. W. and Salakhutdinov, R. (2017). “Linguistic Knowledge as Memory for Recurrent Neural Networks,” ArXiv170302620 Cs.

Ferrández, Ó., Kozareva, Z., Toral, A., Muñoz, R., and Montoyo, A. (2007). "Tackling HAREM’s portuguese named entity recognition task with spanish resources". Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, a primeira avaliação conjunta na área, chapter 11, pages 137–144. Linguateca.

Fonseca, E. B., Vieira, R. and Vanin, A. A. (2014). "Coreference Resolution in Portuguese: Detecting Person, Location and Organization,” Learn. Nonlinear Models, vol. 12, pp. 86–97.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings ofthe eighteenth international conference on machine learning, pages 282–289

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016). Neural Architectures for Named Entity Recognition.

Martschat, S. and Strube, M. (2015). "Latent structures for coreference resolution." Transactions of the Association for Computational Linguistics, 3, vol. 3, no. 1, pp. 405–418.

Nadeau, D. and Sekine, S. (2007) “A survey of named entity recognition and classification”. Journal Linguisticae Investigationes, National Research Council, vol. 30, pp. 3-26

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J. and Ward, R. (2016). “Deep Sentence Embedding Using Long Short-term Memory Networks: Analysis and Application to Information Retrieval”. IEEEACM Trans Audio Speech Lang Proc, vol. 24, no. 4, pp. 694–707.

Pires, R. A. (2017). FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO Named entity extraction from Portuguese web text.

Ponomareva, N., Rosso, P., Pla, F. and Molina, A. (2007, September). Conditional random fields vs. hidden markov models in a biomedical named entity recognition task. In Proc. of Int. Conf. Recent Advances in Natural Language Processing, RANLP (Vol. 479), p. pages 479–483

Ristad, E. S. and Yianilos, P. N. (1998). “Learning string-edit distance”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 5, pp. 522–532.

Ronan, C. and Weston, J. (2008). A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. TUGboat, 16(3), 233–243.

Sang, E. F., and De Meulder, F. (2003). “Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition,” in Proceedings of the Seventh Conference on Natural Language

Santos, D. and Cardoso, N. (2008) “Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área,

Teixeira, J., Sarmento, L. and Oliveira, E. (2011, October). "A bootstrapping approach for training a ner with conditional random fields". In Portuguese Conference on Artificial Intelligence,pages 664–678. Springer Berlin Heidelberg.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L. and El-Bachouti, M. (2013). Ontonotes release 5.0 ldc2013t19. Linguistic Data Consortium, Philadelphia, PA.

Wiseman, S. J., Rush, A. M., Shieber, S. M. and Weston, J., (2015) Learning anaphoricity and antecedent ranking features for coreference resolution. Association for Computational Linguistics.