Automatic Extraction of Entities and Relation from Legal Documents

Judith Jeyafreeda Andrew[1]
GREYC , Campus 2 UniCaen,
Bâtiment F
6 Boulevard Maréchal Juin , 14000 Caen
judithjeyafreeda@gmail.com

Xavier Tannier[1]
Sorbonne Université,
Inserm, LIMICS, Paris, France
xavier.tannier@
sorbonne-universite.fr

Abstract

In recent years, the journalists and computer sciences speak to each other to identify useful technologies which would help them in extracting useful information. This is called "computational Journalism". In this paper, we present a method that will enable the journalists to automatically identifies and annotates entities such as names of people, organizations, role and functions of people in legal documents; the relationship between these entities are also explored. The system uses a combination of both statistical and rule based technique. The statistical method used is Conditional Random Fields and for the rule based technique, document and language specific regular expressions are used.

1 Introduction

Everyday there are a number of legal documents that are being recorded and made available as text documents. In this paper, we present a system that automatically identifies named entities and the relationships between various entities within a dataset of certain type of legal documents which contains information about people investing in property. This helps journalists to identify some useful information - information like the name of the person investing and company invested in. We propose a hybrid method to automatically detect different types of relationship after identifying the entities within the corpus. We follow a combination of statistical and rule based techniques to achieve the goal.

The objective of this project therefore are:

- To identify and classify the entities within each of the text documents
- To identify the relationships between the entities

To achieve the objectives, we present a hybrid system which explores a combination of two techniques for Named Entity recognition (a statistical approach using Conditional Random Fields (CRF) and rule based techniques) and produces a graph with all entities and their relationships, in the perspective of a investigative journalism use.

2 Data

The data used in this project is a corpus taken from the so-called "Luxembourg" corpus. This publicly available legal register contains information about people and companies who are investing money or property in the state of Luxembourg. Most of the documents are written in French, and we only worked on this language.

Some of the data set has been annotated manually with the help of the brat tool (Stenetorp et al., 2012) for the different classes and the relationship between the classes by our journalist partners. The annotations have been done manually for 35 documents which can be used as a training set to develop a model.

2.1 Entities

The classes used for classification of the entities are as follows:

- PERSONNE represents the name of the person
- NOM represents the first name of the person
- ADDRESS represents the address of the organization

[1]The work was done while the authors were affiliated with LIMSI, CNRS and Univ. Paris-Sud.

Proceedings of the Seventh Named Entities Workshop, pages 1–8
Melbourne, Australia, July 20, 2018.

- SOCIETE_PRINCIPALE represents the name of the main company participating in the transaction
- SOCIETE_SECONDAIRE represents the name of the secondary companies participating in the transaction
- ROLE represents the role of the identified person or company in the transaction
- FONCTION is the function or position held by the identified person in the transaction
- TYPE_SOCIÉTÉ is the type of the companies identified

2.2 Relations

The relationships between the entities are classified as follows:

- 'PERSONNE_FONCTION" is the relationship between the class "PERSONNE" and the class "FONCTION"
- "PERSONNE_ROLE" is the relationship between the class "PERSONNE" and the class "ROLE"
- "SOCIÉTÉ_ROLE" is the relationship between the class "SOCIÉTÉ" and the class "ROLE"
- "SOCIÉTÉ_TYPE" is the relationship between the class "SOCIÉTÉ" and the class "TYPE_SOCIÉTÉ"

2.3 Structure of the corpus

The structure and language of legal documents are more rigid than free text. When the persons and companies are identified, then the other classes appear in the same sentence and can be identified by only a few specific expressions. Below are few examples, the translation in English are given in the "[]".

- "Ensuite les souscripteurs prédésignés, représentés par Me Catherine Dessoy, prénommée, en vertu des procurations susvantées" ["Then the underwriters, represented by Catherine Dessoy, prenamed, under the aforementioned powers of attorney"], where "représentés par" is the ROLE and "Me Catherine Dessoy" is the PERSONNE.
- "Par-devant Maître Blanche Moutrier, notaire de résidence à Esch-sur-Alzette."["Before Maître Blanche Moutrier, notary of residence in Esch-sur-Alzette."], where "Maître Blanche Moutrier" is the PERSONNE and "notaire" is the FONCTION.
- "CUBE INVEST S.A.-SPF, une société de gestion de patrimoine familial, en abrégé SPF, sous forme d'une société anonyme" ["CUBE INVEST S.A.-SPF, a family wealth management company, in abbreviated SPF, in the form of anonymous company"], where "CUBE INVEST S.A.-SPF" is the SOCIÉTÉ and "société de gestion de patrimoine familial" is the TYPE_SOCIÉTÉ.

Because of this rigid structure of the legal documents, rule-based techniques will be able to identify some of the entities. However, the basic classes PERSONNE and SOCIÉTÉ have to be identified first in order to take advantage of this rigid structure. Figure 1 shows an example of an annotated document as seen by the BRAT visualization tool emphasizing on the structure of the legal documents.

2.4 Training and Test data sets

The data is divided into training and test set. The training set is a set of corpus consisting of 35 text files and test set is a collection of 21 text documents. The method has been trained and tested on this small corpus, however it is developed with the scope of being able to build a graph with all the documents available in the Luxembourg register. This mounts up to data between the years 2002 to 2016, containing about 2,041,111 text documents. For this reason, the training documents have been taken randomly from the entire collection.

3 Related Work

3.1 Conditional Random Fields

Conditional Random Field (CRF (Lafferty, 2001)) is a sequence modeling technique belonging to the class of statistical modeling methods. It is often used in labeling and parsing sequential data. A CRF has a single exponential model for the joint probability of the entire sequence of labels given the observation sequence. (Sutton and McCallum, 2012) gives a detailed tutorial on Condition Random fields. Since the CRF model is conditional, dependencies among the input variables x

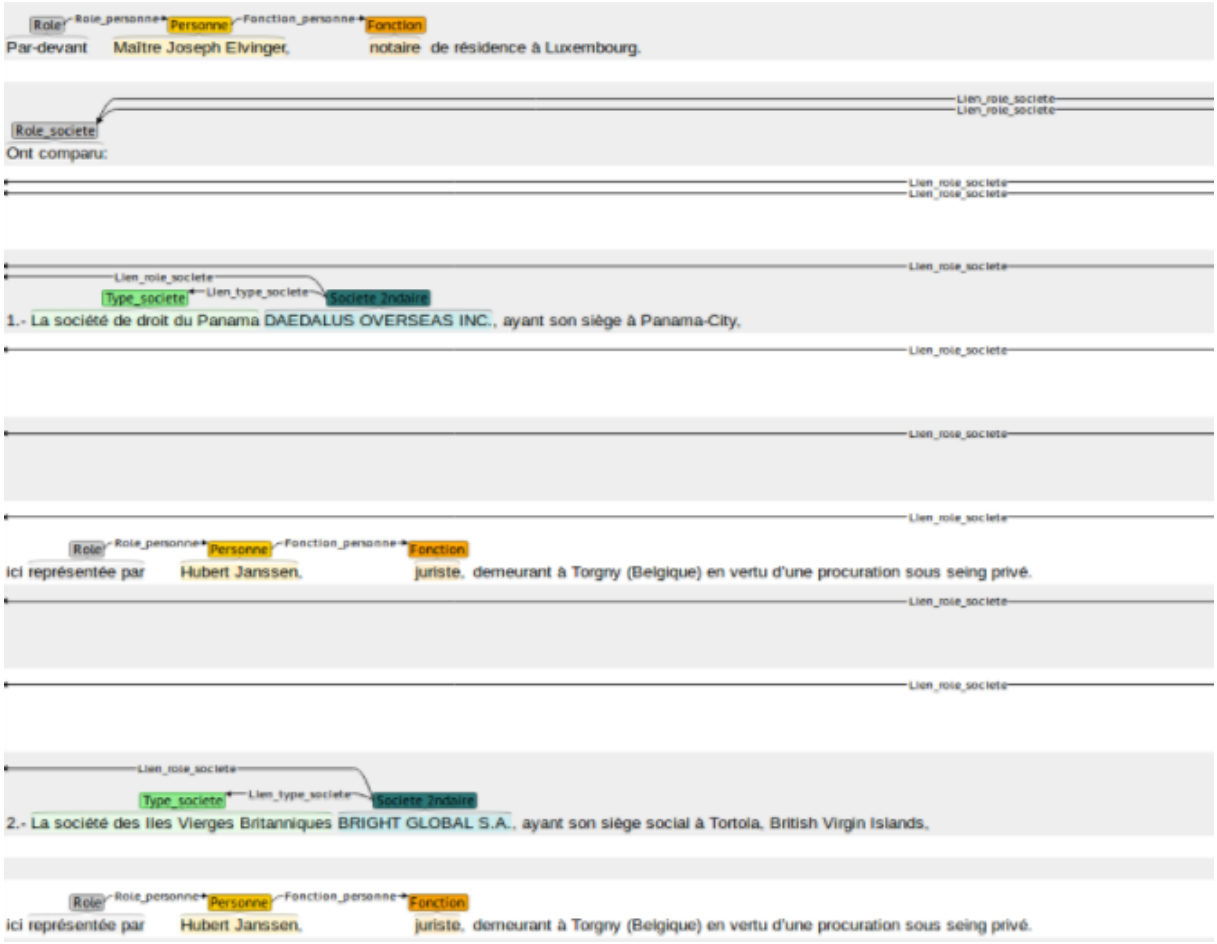

Figure 1: Annotated document presented with the BRAT visualization tool

do not need to be explicitly represented. This allows CRF to be used widely in Natural Language Processing. (Sutton and McCallum, 2012) also suggests that some of the useful features that could be used in Natural Language Processing are capitalization, word bigrams, neighboring words etc. In this work, word bigrams and capitalization have been used extensively.

3.1.1 Conditional Random Fields for Entity Recognition

There have been quite a lot of work done with respect to entity recognition and classification using CRF.

(N.V et al., 2010) describes the use conditional Random Fields for Entity Recognition in geological text. (McCallum and Li, 2003) presents a named entity recognition technique with conditional random fields, where web enhanced lexicons are used for feature induction. (Ghamrawi and McCallum, 2005) present the multi-label classification of corpora using classification. Multi-label classification is a task of assigning an object simultaneously to one or multiple classes. (Ghamrawi and McCallum, 2005) present two graphical models for multi-label classification, namely the Collective Multi-Label classifier and the Collective Multi-Label with Features classifier. CRFs have better performances than many other techniques. (Li et al., 2008) compares SVM with CRF for named entity recognition with clinical data and concludes that CRF outperforms SVM.

4 Approach

The approach used is a combination of statistical approach (CRF) and the rule based technique.

4.1 Process

In order to annotate the corpus with the entities and the relationship, the work uses two techniques which are conditional random fields and rules based on regular expressions. Conditional Random Fields (CRF) is used to annotate the document only for the classes "PERSONNE", "SOCIÉTÉ", "NOM","ADDRESS". These classes are the basic classes and therefore they have to be identified first. Moreover, we only expect the other classes to appear in the same sentence as a "PERSONNE" or a SOCIÉTÉ or a "NOM" or a "ADDRESS". Therefore, identifying these classes will be the first and basic step. For the other classes, a rule based technique are used. 2 shows the process flow used for the annotation of text. The rules are written in such a way that they identify the other classes and their relation with the main classes ("PERSONNE" and "SOCIÉTÉ").

5 Implementation

5.1 Conditional Random Fields (CRF)

In order to annotate the document for the base classes ("PERSONNE", "SOCIÉTÉ", "ADDRESS", "NOM"), Conditional Random Fields are used. The system uses the wapiti toolkit (Lavergne et al., 2010) to train the CRF.

In order to use wapiti, the training set and the test set are converted into the BIO format. Figure 3 shows how the wapiti tool works in order to train and test using CRF.

In order to use conditional random fields, one has to create a pattern file with which CRF can be trained. A pattern file defines some features that are going to be used by the wapiti.

5.2 Regular Expressions for entity recognition and relationship

In order to identify the other classes (ROLE, FONCTION, TYPE_SOCIÉTÉ), regular expressions are used. The rules are written such that once the entities are identified the relationship can be established with the same rule. This is done by writing the rules using the relationship itself. For example, if there is a "PERSONNE" in a sentence, then the sentence should have a "ROLE" and a "FONCTION" for the identified person. This suggests that there exist a relationship between the person and his/her "ROLE" and "FONCTION". Therefore, the entities "ROLE" and "FONCTION" should occur somewhere close to the entity "PERSONNE". This rule-based system is established with the help of the GATE tool (Cunningham et al., 2011) and the rules are written as JAPE grammar(Thakker et al., 2009)

5.2.1 Formation of JAPE rules for the various classes

Class "FONCTION" The class "FONCTION" is the job of the person in question. The GATE gazetteer is used to annotate the function of a person. GATE gazetteer does not have a dictionary for the function of a person. Therefore, a dictionary is created with all the words that could be the function of a person. This dictionary has been cre-

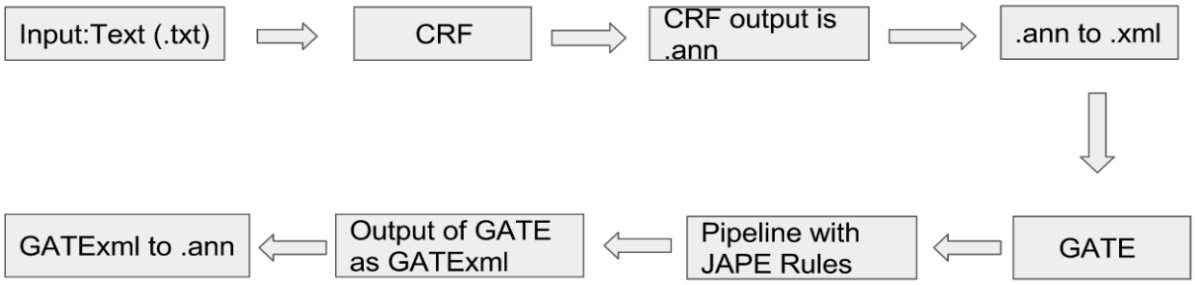

Figure 2: Process Flow for identifying entities and relationships

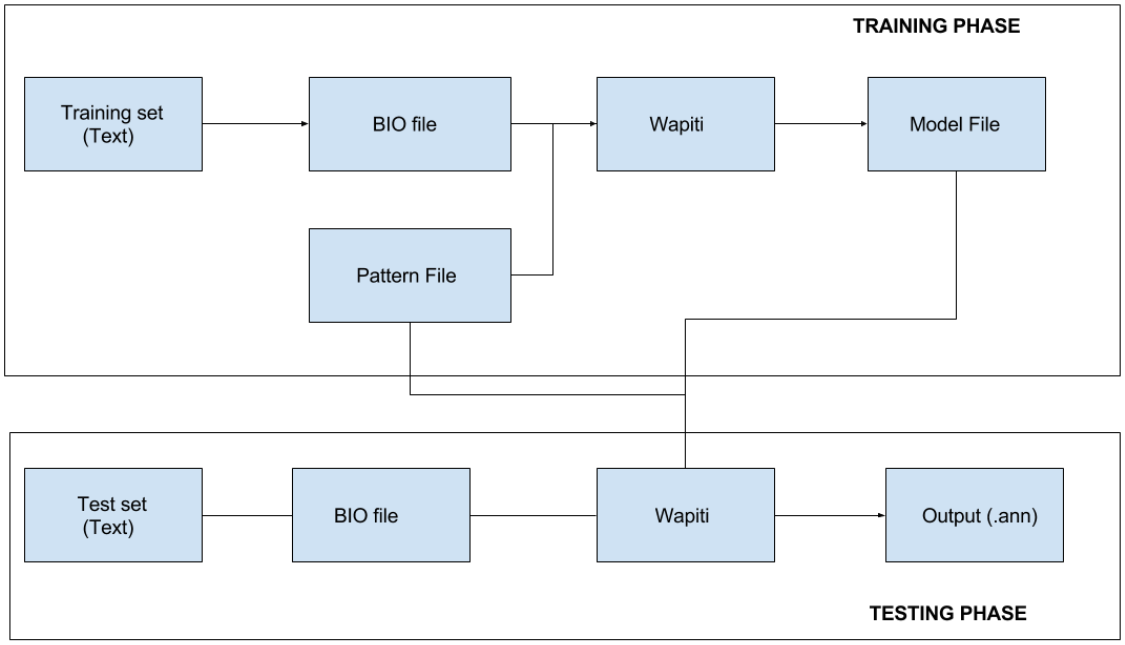

Figure 3: The process flow as followed by wapiti

ated with about 500 words and added to the GATE gazetteer. The rule for the class "FONCTION" was developed as per the structure of all the documents in the training set, where it was identified that the class "FONCTION" appears with the same sentence as the class "PERSONNE". Thus using the gazetteer and the class "PERSONNE", the class "FONCTION" can be annotated and the relation 'PERSONNE_FONCTION" is drawn. For example, "Par-devant Maître Blanche Moutrier, notaire de résidence à Esch-sur-Alzette." ["Before Maître Blanche Moutrier, notary of residence in Esch-sur-Alzette."]. In this sentence, "Maître Blanche Moutrier" is the class "PERSONNE" and this is followed by the "notaire" which is the "FONCTION" of "Maître Blanche Moutrier".

class "ROLE" The "ROLE" of a "PERSONNE" almost always occur in the same sentence as the class "PERSONNE". It could occur after or before "PERSONNE" "ROLE" could also be associated with the class "SOCIÉTÉ" as well. In this case, the class "ROLE" occurs in the same sentence as the "SOCIÉTÉ". This leads to the "PERSONNE_ROLE" and "SOCIÉTÉ_ROLE" relation.

For example: "Pardevant Maître Henri Hellinckx, notaire de résidence à Luxembourg." ["Late Maître Henri Hellinckx, notary residing in Luxembourg."] In the above sentence, "Pardevant" is the "ROLE" played by "Maître Henri Hellinckx" and the "FONCTION" is "notaire". Therefore the rule is to identify the sequence "PERSONNE" "FONCTION" and identify the word before the sequence as "ROLE".

Another example: "Ont comparu: 1.- La sociètè de droit du Panama DAEDALUS OVERSEAS INC., ayant son siège à Panama-City" ["Appeared: 1.- The company of law of Panama DAEDALUS OVERSEAS INC., Having its headquarters in Panama-City"], where "Ont comparu" is the "ROLE" and "DAEDALUS OVERSEAS INC." is the "SOCIÉTÉ" . Here the "ROLE" is followed by tokens like punctuations and numbers before the "SOCIÉTÉ" which also have to be incorporated with the rules.

Another example: "Les parts sociales ont ètè souscrites par LUXEMBOURG CORPORATION COMPANY S.A., prèqualifièe, qui est l'associèe unique de la socié té " ["The shares have been subscribed by LUXEMBOURG CORPORATION COMPANY S.A., prequalified, which is the sole partner of the company."], where " parts sociales ont été souscrites" and "l'associèe unique de la société ." are both "ROLE" of the "SOCIÉTÉ" "LUXEMBOURG CORPORATION COMPANY S.A." This state of having two roles is handled with a different rule as well.

Therefore, in order to help handle all these different situations, multiple different rules are used. A total of 20 different JAPE rules has been written to annotate all the roles in all the different situations. This count includes identifying the roles of the société as well.

class "TYPE_SOCIÉTÉ" The class "TYPE_SOCIÉTÉ" tells about the type of the "SOCIÉTÉ". Therefore the type has to be occurring in the same sentence as the "SOCIÉTÉ". It is also identified that all the texts in the training set had the type of the société in the same sentence as the SOCIÉTÉ. Also, the type of the société always starts with the word "société" followed by a type. This then leads to the relationship of "SOCIÉTÉ_TYPE". Example: "S'est réunie l'Assemblée Générale Extraordinaire des associé s de la société à responsabilité limité e thermo haus, S.à r.l., ayant son siè ge social à L-6940 Niederanven, 141, route de Trèves, inscrite au Registre du Commerce et des Socié tés à Luxembourg, section B sous le numéro 74.172, constituée suivant acte reçu par Maître Alex Weber, notaire de résidence à Bascharage, en date du 2 février 2000, publié au Mémorial C de 2000, page 16652." ["The Extraordinary General Assembly of the associates of the limited liability company, S.à rl, having its if it is located at L-6940 Niederanven, 141, route de Trvesves, entered in the Register of Commerce and Companies in Luxembourg, section B under number 74.172, incorporated according to the deed of the Court, given to Alex Weber, notary residing at Bascharage, on February 2, 2000, published in the Mémorial C of 2000, page 16652."]. In the above sentence, "thermo haus, S.à r.l." is the SOCIÉTÉ and "la société à responsabilité limitée" is the TYPE_SOCIÉTÉ.

5.2.2 The GATE pipeline

The JAPE rules are incorporated with the other inbuilt modules of the GATE tool to create a pipeline. A GATE pipeline with modules for tokenization , POS tagging along with the JAPE rules is used to annotate the document for the other classes.

Mode	True Positive	False Positive	False Negative	Precision	Recall
Exact	318	61	0	83.91%	100%
Partial	348	30	0	92.08%	100%

Table 1: Results of brat evaluation tool on the training set.

Mode	True Positive	False Positive	False Negative	Precision	Recall
Exact	81	4	88	95.29%	47.93%
Partial	191	9	12	95.50%	94.09%

Table 2: Results of brat evaluation tool on the test set.

6 Evaluation

For the evaluation of annotations, the brat evaluation tool is used (Stenetorp et al., 2012). The comparisons can be done in two ways: either by comparing the file for exact matches or by partial matches. By exact matches we mean that the offset have to be exactly matched between the two files. By partial matches, we mean that even if the offsets do not match perfectly, partial annotations are also considered to be correct.

7 Results

The results are shown in Tables 1 and 2. Table 1 corresponds to the results from the training set and the table 2 corresponds to the results from the test test. The results depend on both the processes - the CRF and the rule based technique. The performance of the CRF is with an error rate of 3.12%.

The low recall value for the exact matches of the test set as compared with the training set is due the tailoring of the rules. While training, the data set has been referred to at many times to come up with expressions that will help in retrieving all the possible instances of every annotations. However, while the same rules have been run on test data which has not been seen before hand, it is noted that there requires many more rules that need to added to the already existing rules to improve the recall value.

However it has to be noted that the recall value is quite high with the partial matches. For example: instead of annotating "ici représenté par", it annotates "représenté par". This is not totally wrong. Considering the knowledge base, this annotation is still useful. Though it is not the exact same annotation as in the manual annotation, it is still considered valid.

Thus considering the results of partial annotations only, this method proves to be quite efficient in annotating the files from the "Luxembourg" register.

As indicated above, the process has been developed over a small set of data, but the process can be run over huge volumes of data. The total amount of documents tested are 2,041,111. The number of relations found in these documents are 3,026,560. However, since these data have no manual annotations, no evaluation was performed on this set of data.

References

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. Text Processing with GATE (Version 6).

Nadia Ghamrawi and Andrew McCallum. 2005. Collective multi-label classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 195–200, New York, NY, USA. ACM.

John Lafferty. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289. Morgan Kaufmann.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical Very Large Scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513, Uppsala, Sweden. Association for Computational Linguistics.

Dingcheng Li, Karin Kipper-Schuler, and Guergana Savova. 2008. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 94–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 188–191.

Sobhana N.V, Pabitra Mitra, and S.K. Ghosh. 2010. Article: Conditional random field based named entity recognition in geological text. *International Journal of Computer Applications*, 1(3):119–125. Published By Foundation of Computer Science.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Charles Sutton and Andrew McCallum. 2012. An introduction to conditional random fields. *Found. Trends Mach. Learn.*, 4(4):267–373.

Dhaval Thakker, Taha Osman, and Phil Lakin. 2009. GATE JAPE grammar tutorial.