Named Entity Disambiguation in Biomedical Literature

Matthew Gittings

Master of Science

School of Computer Science

McGill University

Montréal, Québec

August 15, 2018.

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

DEDICATION

This thesis is dedicated to my parents, whose wisdom and support I am incredibly grateful for.

ACKNOWLEDGEMENTS

I would like to sincerely thank my supervisor, Professor Derek Ruths, for his guidance, knowledge, and mentorship. I would also like to thank Professor Nicholas King for his active engagement and insights.

ABSTRACT

The presence of conflict of interest in biomedical literature is a space of increasing interest in public health, due to concerns of entities driving research outcomes for their benefit. However, conflict of interest, and the potential influence that it exerts on research outcomes, remains a difficult phenomenon to capture and investigate on a large scale. In this thesis, we present and characterize a dataset of researchers and their engagements with entities, extracted from article conflict of interest disclosures, and present methods to disambiguate the textual references to real-world entities.

We first present a classification model for linking entity mentions locally in the same article, and then move on to disambiguating entity mentions globally from across articles. We then tune and apply these methods to our entire dataset, yielding a large researcher-entity network, that can be leveraged in the study of conflict of interest. Finally, we present one such analysis where we characterize industry engagement through the lens of researcher influence.

We find that the task of local entity linking can be performed with high accuracy, and that global entity disambiguation, though more involved and requiring more tuning, can be performed to a high degree as well. We also present evidence from a network analysis that shows that corporations target researchers with higher levels of influence in their fields than other entity types of government organizations and foundations.

ABRÉGÉ

La présence de conflits d'intérêts dans la littérature biomédicale est un champ d'intérêt croissant en santé publique, en raison des préoccupations en lien avec les entités qui génèrent des résultats de recherche à leur avantage. Néanmoins, les conflits d'intérêts et l'influence potentielle qu'ils exercent sur les résultats de recherche demeurent un phénomène difficile à saisir et à étudier à grande échelle. Dans cette thèse, nous évaluerons un ensemble de données de chercheurs et leurs engagements avec des entités, extraits de la divulgation d'un conflit d'intérêts d'article, et des méthodes actuelles pour déchiffrer les références textuelles à des entités du monde réel.

Nous présentons d'abord un modèle de classification pour lier les mentions d'entités localement dans le même article, puis nous passerons à la démystification des mentions d'entités à travers les articles. Nous ajusterons ensuite et appliquerons ces méthodes à l'ensemble de notre base de données, ce qui donnera un vaste réseau d'entités, qui peut être exploité dans l'étude des conflits d'intérêts. Enfin, nous présenterons cette analyse où nous caractérisons l'engagement de l'industrie à travers la lentille de l'influence des chercheurs.

Nous énoncerons que la liaison d'entités locales peut être effectuée avec une grande précision, et que la désambiguïsation d'entité globale, bien que plus impliquée et nécessitant plus de réglage, peut également être réalisée à un haut niveau. Nous présentons aussi des preuves provenant d'une analyse de réseaux qui montre que les

entreprises ciblent les chercheurs ayant des niveaux d'influence plus élevés dans leurs domaines que les autres types d'organisations gouvernementales et de fondations.

CONTRIBUTIONS OF AUTHORS

This thesis builds off previous work in conflict of interest, pursued by Professor Derek Ruths, Professor Nicholas King, and Sarah Berry. This includes the conflict of interest dataset and relationship definitions presented in Chapters 2 and 3, as well as parts of the entity type classification work presented in Chapter 2, and used throughout this thesis. The rest of the work presented is my own, though with consultation with the rest of the team.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

CHAPTER 1
Introduction

1.1 Introduction

There is widespread concern regarding conflict of interest in biomedical literature, and the potential influence that various entities exert over research outcomes in the medical field, through engagement with researchers. Such engagement manifests in the form of grant funding, honorariums, salary, and various other support mechanisms, and could be leveraged to influence research outcomes in favour of the entities. However, this phenomenon remains difficult to quantify and investigate. Research on this topic is primarily limited by the absence of large-scale datasets, that capture the engagement between authors, articles, studies and entities in the field of biomedicine.

In recent years, many journals have established processes requiring that researchers declare any engagements with entities that may constitute a conflict of interest. The International Committee of Medical Journal Editors (ICMJE) distributes a conflict of interest form that is used by many medical journals, to be submitted alongside publications. This form aims to capture full disclosure of any engagements that may constitute conflict of interest for an author, and creates a public record of engagements between authors and entities in this domain. While engagement does not inherently constitute a conflict of interest or influence, it is a prerequisite for both. Making these records public serves to inform educators, policy makers, and the public, and enable informed judgments of the research results and outcomes.

While publicly available, these records are distributed across different journals, associated with individual studies, and in plain text. This format is not conducive to evaluating or quantifying industry engagement in biomedicine, or modelling author-entity engagements, as there is no central knowledge base that documents all relationships between researchers and entities. To create a snapshot of industry engagement, individual conflict of interest statements from across publications must be inspected and coded, to identify various entities and their relationships with the authors. The task of identifying and extracting entities from the conflict of interest disclosure statements is known formally as *named entity recognition (NER).* This task involves "identifying the names of all the people, organizations, and geographic locations in a text" [10]. In this context, the named entities of interest are the people (researchers), and any organizations (entities) they declare engagements with. The task of identifying the relations between named entities is known as *relation extraction.*

However, identifying entity-author engagement pairs is not enough to perform any meaningful analysis. While named entity recognition can identify pairs of authors and any entities that they are engaged with, it only does so at a local (study or article) level. It is not only possible, but highly likely, that the same authors and entities appear multiple times across various publications, and with various resolutions. Thus, the authors and entities identified via named entity recognition must also be *disambiguated* against other named entities, so that the engagement network accurately represents real-world interactions. This process of disambiguating named entities is known formally as *named entity disambiguation (NED)*, and is a main focus of this work. The objective is to resolve coreferent entity mentions in conflict of interest statements, such that all entity mentions that represent the same real-world entities are identified and grouped.

Our motivation is to create a high resolution engagement network of researchers, articles, and entities, that accurately captures and depicts real-world interactions. Such a network will provide descriptive statistics, enable high-impact network analysis to identify entity engagement strategies with authors, identify communities of researchers and entities, and other analyses. It may also eventually enable recommendations for disclosure practices in an effort to inhibit conflict of interest.

This thesis presents methods and analyses for linking and disambiguating entity mentions extracted from biomedical conflict of interest statements. The results are then used to produce an engagement network of entities and researchers, and a sample analysis of this network is presented.

1.2 Outline

The following chapters of this thesis are outlined below:

Chapter 2 outlines prior work in conflict of interest, named entity recognition, and named entity disambiguation, as well as the terminology used throughout this thesis. This is followed by prior work on the evaluation metrics, classification models, validation techniques, and similarity measures used in the entity disambiguation process.

Chapter 3 defines the exact entity disambiguation problem to be solved, and describes the data used in this work. It then outlines how the previously depicted methods and models are leveraged to disambiguate entities at a local (intra-article) and global (inter-article) level.

Chapter 4 presents the local and global disambiguation results, using various evaluation metrics.

Chapter 5 presents a discussion on the local and global disambiguation results, as well as discussion on how the evaluation metrics are leveraged to produce final entity clusters of optimal resolution.

Chapter 6 presents an application of how the disambiguated dataset can be used to perform meaningful analysis in the field of conflict of interest, where we investigate engagement patterns by each entity type with researchers.

Finally, Chapter 7 provides a final discussion and conclusion of the results, and on the overall contributions of this work.

CHAPTER 2
Related Works and Background Information

In this chapter, we define the background information and context that motivates this research, as well as the terminology that will be used throughout this thesis. We also present necessary background work on the topics, models, and methods that are leveraged throughout.

2.1 Conflict of Interest

There is much interest in quantifying the impact of engagement between pharmaceutical and biotechnology companies and researchers in the field of biomedicine [21]. The implications of conflict of interest may have a drastic impact on the integrity and validity of scientific results, and thus it is important to be aware of industry engagement when considering scientific results. It is also then important to define exactly what constitutes a conflict of interest. For this, we borrow the following definition from Dennis Thompson [27]:

> A conflict of interest is a set of conditions in which professional judgment concerning a primary interest (such as a patient's welfare or the validity of research) tends to be unduly influenced by a secondary interest (such as financial gain).

Prior work of this type has explored conflict of interest in biomedical research, and found that corporate support of studies tends to yield favourable results for the supporting entity [22]. Furthermore, this influence may not solely manifest in research and clinical trials - it is also common for industry to engage with physicians directly [13], however in this work we focus on the former.

2.2 Named Entity Recognition

Named entity recognition (NER) is the process of identifying and annotating proper nouns in text. The three most common labels for these classes are person, organization, and location [19]. These annotations are denoted with "PER", "ORG", and "LOC" respectively. For named entities that span multiple word tokens, there are two common naming conventions: BIO, and BILOU. BIO stands for "**B**eginning, **I**nside, **O**utside". Any token that begins a named entity is labelled with a B, any proceeding token that remains part of the same entity is labelled with an I, and any token that is not part of a named entity is represented with an O. The BILOU scheme is slightly more descriptive, and stands for "**B**eginning, **I**nside, **L**ast, **U**nit, **O**utside". The additional L and U labels indicate the final token of a named entity token sequence, and any single-token named entity respectively [24]. For example, consider the following excerpt from a conflict of interest statement:

$\underbrace{AUTHOR\_1}_{\text{E-PER}}$ reports having received research grant support and lecture fees from $\underbrace{\textit{Bristol-Myers Squibb (BMS)}}_{\text{E-ORG}}$ and $\underbrace{Merck}_{\text{E-ORG}}$.

Under the BIO and BILOU schemes, we get the representations of this sequence shown in Table 2–1:

Table 2–1: BIO and BILOU representations of entities.

Token	**BIO**	**BILOU**
AUTHOR_1	B-PER	U-PER
reports	O	O
having	O	O
received	O	O
research	O	O
grant	O	O
support	O	O
and	O	O
lecture	O	O
fees	O	O
from	O	O
Bristol-Myers	B-ORG	B-ORG
Squibb	I-ORG	L-ORG
(BMS)	B-ORG	U-ORG
and	O	O
Merck	B-ORG	U-ORG
.	O	O

For the given context of conflict of interest in biomedical literature, we are ultimately concerned with the E-PER and E-ORG classes, as we model relations between authors and entities. In Section 2.5, we further elaborate on these classes as they pertain to this work.

2.3 Named Entity Disambiguation

Named entity disambiguation (NED) is the process of determining the real-world entity corresponding to an *entity mention* in the text. In our context, it is the process of determining which entity mentions across the entire sample refer to the same real world entities, and grouping these into a single cluster. For example, the following entities from the data occur across different conflict of interest statements, but should be clustered together for analysis purposes:

1. USAID
2. U.S. Agency for International Development
3. U.S. Agency for International Development (USAID) Emerging Pandemic Threats Program
4. USAID Victims of Torture Fund
5. USAID National M&E Support Programme
6. US AID

Each of these strings is a different way of representing the same overarching entity. While they may differ in terms of resolution, in the context of this work and for the data resolution we desire, collapsing these entity mentions is desirable. We are concerned with analyzing large-scale patterns of engagement between entities and researchers, and thus we are interested in macro level entities. Another factor contributing to this need is the quantity of data. If each of these entity mentions were given their own node in a network, our relationship sample size for each entity decreases, because the network becomes more sparse. This in turn affects network analysis outcomes.

2.4 Record Linkage

This named entity disambiguation process is very similar conceptually with *record linkage*, which is the task of identifying records in data that correspond to the same entity, and associating them [4]. There are two broad categories of record linkage: deterministic and probabilistic. Deterministic record linkage uses rule-based linking criteria, where records are linked if a given set of conditions are met, in the form of a predicate. In this approach, a similarity score is computed between various features from two candidate records, and then rules are established to assign the match or non-match label. This differs from probabilistic record linkage, where a binary classification model is trained using the features of two candidate records, to estimate the probability of a link. A link is then assigned if the estimated probability exceeds a threshold.

2.5 Terminology

The Automated Content Extraction (ACE) program defines challenges for detection of entities and relations in natural language, and the Linguistic Data Consortium (LDC) provides well-adopted annotation guidelines for these tasks. They define an *entity* as "an object or set of objects in the world". Furthermore, they define an *entity mention* as "a textual reference to an entity" [5].

2.5.1 Entity Types

The LDC also defines 7 types of entities: "Person", "Organization", "Facility", "Location", "Geo-political Entity", "Vehicle", "Weapon" [5], of which two, organizations and persons, are relevant to this work, as we are interested in relations between authors and entities. Organizations can be broken down into 5 sub-types: "Government", "Commercial", "Educational", "Non-profit", and "Other". In the context of biomedical literature however, we adapt these sub-types to better suit our purposes:

1. **Government**: Any organization that is part of a national, federal, provincial, state, municipal, or local government.

2. **Corporation**: Any for-profit company that is either publicly or privately owned. There are three sub-types of corporations:
   - **Pharmaceutical or Bio-pharmaceutical**: Any company that provides pharmaceutical or bio-pharmaceutical products or services.
   - **Health/Medical Technology**: Any company that provides health or medical technology products or services..
   - **Combination**: Any company that is deemed to be a combination of either above category.
3. **Academic**: Any organization that is engaged in higher education or research, such as a university, college, or academic institution.
4. **Foundation**: Any non-profit organization that makes grants to unrelated entities for scientific, educational, or other charitable purposes.
5. **Hospital**: An organization that provides medical services to patients.
6. **Professional Association**: An organization that represents practitioners of a given occupation, and oversees certifications in this field.
7. **Partnership**: Any entity that is deemed to be a partnership between two entities of differing types defined above.

Each entity mention in the conflict of interest statements is thus assigned one of the entity type labels defined in this section.

2.5.2 Relation Types

The LDC also defines guidelines for relation detection from natural language. The 7 types of entity-entity relations are: “Physical”, “Personal/Social”, “Employment/Membership/Subsidiary (EMP-ORG)”, “Agent-Artifact”, “Person/Organization

Affiliation", "Geopolitical Entity Affiliation", and "Discourse" [6]. Similarly, two of these relations are relevant in the context of this work, the EMP-ORG, and PER/ORG affiliations, as these depict relations between people and organizations. However, the resolution of these sub types is not precise enough for this work. As a result, we again define new sub-types to better suit our context. These author-entity relationship types are shown in Table 2–2. In our data, each entity-author tuple is assigned one of the relation type labels.

Table 2–2: Author-entity relationship types..

#	**Relationship Type**
1	Received research grant directly
2	Speakers' bureau
3	Consultant
4	Equity
5	Employee of
6	Collaborator
7	Scholarship
8	Award
9	Fellowship
10	Honorarium
11	Personal fees
12	Former employee of
13	Founder of entity or organization
14	Holds chair
15	Patent license
16	Named professor
17	Board member
18	Expert testimony
19	Patent
20	Received research materials directly
21	Received research materials indirectly
22	Received research grant funds directly
23	Received research grant funds indirectly
24	Received travel support
25	Salary support
26	Research trial committee member
27	Supported
28	Other/Unspecified

2.6 Evaluation Metrics

Throughout this thesis, several classification models are presented, along with the features used in these models. Here we describe several fundamental concepts and metrics that are leveraged as part of the models, as well as to evaluate them.

2.6.1 Jaccard Index / Jaccard Distance

The Jaccard index of two sets is the size of the intersection of the sets, divided by the size of the union of the sets. This is a measure of similarity between the sets, and is defined mathematically as:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{2.1}$$

This index has a range of $[0, 1]$ and represents the *overlap* of the two sets. Identical sets yield a score of 1, and entirely disjoint sets yield a score of 0.

Similarly, there is a Jaccard *distance* between two sets. It is a measure of dissimilarity between the two, and is defined mathematically as:

$$\text{Distance}_{\text{Jaccard}}(A, B) = 1 - \text{Jaccard}(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{2.2}$$

This index also has a range of $[0, 1]$, and is the size of the symmetric difference of the sets, divided by the size of the union.

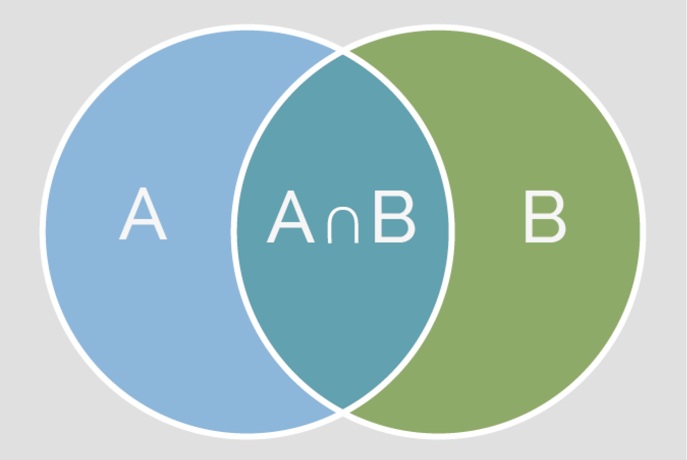

Figure 2–1: Jaccard index of two sets.

Figure 2–1 visualizes the union, intersection, and symmetric difference of sets A and B. The Jaccard index measures the relative size of the intersection, $|A \cap B|$, to the union $|A \cup B|$, whereas the Jaccard distance measures the relative size of the symmetric difference, $|A \cup B| - |A \cap B|$, to the union.

2.6.2 Accuracy, Precision, Recall, and F-Score

Accuracy, precision, recall, and F-Score are commonly used measures for evaluating predictions in binary classification. Accuracy is the number of correct predictions, divided by the total number of predictions made. It is thus a metric for the overall performance of the classifier. It is defined mathematically as:

$$\text{Accuracy} = \frac{\text{T.P.} + \text{T.N.}}{\text{T.P} + \text{T.N.} + \text{F.P.} + \text{F.N.}} \tag{2.3}$$

where T.P. denotes the *true positive* predictions, T.N. denotes the *true negative* predictions, F.P. denotes the *false positive* predictions, and F.N. denotes the *false negative* predictions. However, for unbalanced classes, accuracy is often a misleading performance metric, as simply always predicting the dominant class often yields a high accuracy.

As a result, *precision* and *recall* are often used in place of accuracy. Precision is the number of correct predictions of a given class, divided by the total number of predictions of that class. It represents the fraction of times the classifier was correct when it predicted the positive class. Recall is the number of correct predictions of a given class, divided by the true number of that given class. It represents the fraction of times the classifier predicted a positive class to belong to the positive class.

$$\text{Precision} = \frac{\text{T.P.}}{\text{T.P.} + \text{F.P.}} \tag{2.4}$$

$$\text{Recall} = \frac{\text{T.P.}}{\text{T.P.} + \text{F.N.}} \tag{2.5}$$

Finally, the F-score is a combination of both precision and recall of a classification model, where the importance of precision or recall is weighted through the parameter β:

$$\text{F}_\beta = \frac{(\beta^2 + 1)(\text{precision} \times \text{recall})}{(\beta^2 \times \text{precision}) + \text{recall}} \tag{2.6}$$

When β equals 1, precision and recall are equally weighted, and the F-Score is the harmonic mean of the two:

$$\text{F}_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2.7}$$

Precision-Recall Curve

We plot these evaluation metrics in what is referred to as a *precision-recall curve*, or similarly, the *receiver operating characteristic (ROC)* curve [7]. For each model, we can use a different probability threshold p when assigning a class prediction. Adjusting this threshold has an important impact on precision and recall, and analyzing the relationship between them allows us to tune this parameter accordingly.

2.6.3 Silhouette Coefficient

The Silhouette coefficient is an internal cluster evaluation metric. It measures the cohesion of a clustered item to the cluster in which it is assigned, relative to the other clusters to which it was not assigned. $a(i)$ is defined as the average distance for an item i to all other items within that cluster, and $b(i)$ is the average distance of the same item to its closest other cluster. The Silhouette coefficient is then defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2.8}$$

This metric has a range of [-1, 1], where 1 represents that the item is appropriately clustered, and -1 represents that the item would better belong in the next closest cluster [26].

To gauge the quality of all clusterings, the average Silhouette coefficient across all clusters, referred to as the average Silhouette width, is reported [15]. Finally, note that if the cluster a contains one item, the distance from that item to that cluster is 0.

2.6.4 K-Fold Cross Validation

Cross validation is an evaluation technique used in machine learning to determine how a model will perform on unseen data. In k-fold cross validation, labelled data is split into k subsets, called *folds*. The model is then trained k times, each time on the data from $k - 1$ of the folds, and tested on on the kth fold. We then average any metrics across folds when reporting model performance [20], provided that the variance across folds is acceptable. An extension of k-fold cross validation, called stratified k-fold cross validation, ensures that the class distributions are equal across each fold. In this work, we use cross validation to evaluate the performance of our entity linking classification models. This enables us to quantify how the model will behave when given unseen data that we know the labels for.

2.6.5 Grid Search

Grid search is a common machine learning technique for finding the optimal parameters of a model. It involves defining a set of values for each parameter, and iteratively training the model on each possible configuration of these parameters [23]. After each model is trained, our performance metrics are applied, and the parameters of the best performing model are used as the "optimal" model. We use a grid search to find the optimal parameters for our entity linking models.

2.7 Classification Models

We present here several well researched supervised classification models that are utilized in this work. They are used to predict whether two entity mentions from conflict of interest statements should be linked or not.

2.7.1 Naïve Bayes

Naïve Bayes is a generative binary classification model that is often used as a baseline in classification tasks, due to its simplicity. It estimates a probability that a given input belongs to a given class, $P(y = C_k \mid x)$, using Bayes rule. Naïve Bayes makes the strong assumption that all features are conditionally independent of one another, hence the term *naive*. Under Naïve Bayes, the probability of a given class can be calculated as:

$$\begin{aligned} &P(y = C_k \mid x) = \frac{P(y = C_k)P(x|y = C_k)}{P(x)} \\ &\text{where,} \\ &P(x) = \sum_k P(y = C_k)P(x \mid y = C_k) \end{aligned} \tag{2.9}$$

The probability of a given class, $P(y = C_k)$, is simply the fraction of times that the class appears in the training data, out of the set of all the training samples. The probabilities of the data, $P(x)$ and $P(x|y = C_k)$, are calculated by fitting distributions for each feature for each class, as well as across all classes. The probability can then be determined using these probability distribution functions.

To assign a class to a given datum x, the probability for each class is calculated, and then the maximum likelihood class (the class with the highest probability) is assigned to the given datum. This can also be adjusted by considering various thresholds of p for assigning a datum to a given class.

2.7.2 Logistic Regression

Logistic regression is a discriminative binary classification model, that is also often used as a baseline in classification tasks for its simplicity. It can also be extended for multi-class classification, referred to as multinomial logistic regression. Logistic regression estimates a probability that a given input has a given class, $P(y = c_k|x)$, by modelling the log-odds ratio that an input is one of the two classes. The log-odds decision boundary is equal to:

$$a = \ln \frac{P(x \mid y = 1)P(y = 1)}{P(x \mid y = 0)P(y = 0)} = \ln \frac{P(y = 1 \mid x)}{P(y = 0 \mid x)} \tag{2.10}$$

For logistic regression, we model this decision boundary with a linear function on the features $x_1 \ldots x_n$:

$$a = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n \tag{2.11}$$

We then use a *logistic function*, or *sigmoid function*, to model the probability that x belongs to the positive class:

$$\sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}} \tag{2.12}$$

To find the weights $w_0 \ldots w_n$, we minimize the log-likelihood (cross-entropy) loss function:

$$L(w) = -\sum_{i=1}^{N}[y_i \log \sigma(\mathbf{w}^T\mathbf{x}_i) + (1 - y_i)\log(1 - \sigma(\mathbf{w}^T\mathbf{x}_i))] \tag{2.13}$$

We can then iteratively solve for the weights using gradient descent. First, take the derivative of the loss function with respect to w:

$$\frac{\partial L(w)}{\partial(w)} = -\sum_{i=1}^{N}\mathbf{x}_i(y_i - \sigma(\mathbf{w}^T\mathbf{x}_i)) \tag{2.14}$$

and then use this derivative to update the weights on each iteration:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \alpha_k\sum_{i=1}^{N}\mathbf{x}_i(y_i - \sigma(\mathbf{w}^T\mathbf{x}_i)) \tag{2.15}$$

where α is an arbitrary learning rate.

There are also several regularization techniques that can be applied to logistic regression, by adding a regularization term to the loss function in Equation 2.13.

Lasso regularization adds a term to bound the L_1 norm of the weights, $\lambda \sum_{i=1}^{k} |w_i|$, where λ is the regularization parameter. Ridge regularization adds a term to bound the L_2 norm of the weights, $\lambda \sum_{i=1}^{k} |{w_i}^2|$. In general, L_1 regularization is more robust to outliers in the data, but less stable with respect to variations in the data than L_2 regularization. L_2 regularization also has a unique solution, whereas L_1 regularization can have multiple solutions.

2.7.3 Decision Trees

Decision trees are a class of discriminative model that can be used for both classification and regression. In this section, we focus specifically on Classification and Regression Trees (CART) [1]. A CART decision tree is a binary tree, where each node represents a "split" in the data, yielding two branches. Each leaf node is assigned an output class such that any datum assigned to that that bin is labelled with that class. To train a decision tree, a feature and a value from the data is selected greedily to split on, such that the the cost function is minimized for that split. The same process is then applied recursively on the two branches of the tree, until a stopping condition (maximum depth, minimum leaf node size) has been met. For classification, there are two common cost functions: Gini impurity, and information gain.

Gini impurity represents the probability of being wrong if you were to assign a class to a datum in the set, by randomly sampling from the distribution of classes in the set [2]. It is defined mathematically as:

$$Gini = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i (1 - p_i) \tag{2.16}$$

where J is the number of classes.

Information gain is the difference between the entropy of the data and the conditional entropy of the data, given that a feature has a known value. It is a measure of how much certainty about the data is gained upon learning the value of that feature. It is defined mathematically as:

$$\mathrm{IG(y,\ x)} = H(y) - H(y \mid x) \tag{2.17}$$

where entropy is defined as:

$$H(P) = \sum_{i=1} p_i \log \frac{1}{p_i} \tag{2.18}$$

and conditional entropy is defined as:

$$H(y \mid x) = \sum_{i=1} P(x = i) H(y \mid x = i) \tag{2.19}$$

In this context, by choosing to split the data on a feature such that the Gini impurity is minimized, the "purity" of the data at each leaf node gradually increases until the stopping condition is met. Similarly, by choosing to split the data on features such that the information gain is maximized, the entropy in each subsequent set is minimized.

At each step of the training, each feature and each value for that feature in the data is iterated over. The data is then split into groups for each value on each feature, and the cost function evaluated. The feature-value pair with the optimal cost is then selected as the next node for the tree, and the process repeats until the stopping condition has been met. Due to binary splitting on a specific value for each feature, decision trees are optimal for learning piecewise axis-orthogonal functions. For the same reason, they are less optimal for curvilinear functions.

Given this learning algorithm, it is possible to build a very deep tree such that each datum falls under its own leaf node. However, trees such as this do not generalize well. Because of this, it is necessary to apply stopping criteria. There are two popular methods: maximum tree depth and minimum leaf node size. Enforcing a maximum tree depth will stop training branches of the tree once they grow to a certain depth. With enough data, this prevents highly specific leaf nodes from being generated. Similarly, enforcing a minimum leaf node size will prevent a leaf node from branching once it holds a minimum number of training data points. This prevents these highly specific nodes from occurring in the tree.

2.7.4 Support Vector Classifier (SVC)

A support vector classifier is a maximal margin perceptron used for classification. They work by finding the decision boundary between two classes of data, such that the margin between the nearest training samples of either class to the decision boundary is the largest Euclidean distance possible. The vectors that comprise the maximum margins are called *support vectors*, hence the name of the model. If the classes are not linearly separable, however, the training algorithm will not terminate. In this scenario, there are two options.

The first is to introduce *slack variables*, which relax the constraints on the SVC. They define a new distance such that points on the wrong side of the decision boundary but within this distance are penalized, but still allowed. This enables the model to to find a decision boundary in the data, at the expense that data points can be incorrectly classified.

The second is to implicitly expand the input space to a higher dimension using a kernel function. A kernel function is a function $K(\mathbf{x}_1, \mathbf{x}_2)$ which is a dot product of a feature mapping, $K(\mathbf{x}_1, \mathbf{x}_2) = \phi(\mathbf{x}_1)\dot{\phi}(\mathbf{x}_2)$. Given this function, we replace each $x_i \in \mathbf{x}$ with $\phi(x_i)$. Intuitively, this can be thought of as a "similarity" function between two data points. This process enables different (and potentially better) decision boundaries to be learned. When classifying a new data point, the kernel K calculates the similarity between the point and each of the training points, and predicts the most similar training point. There are many different known kernel functions, including linear, sigmoid, polynomial, and radial basis [12].

2.8 Clustering Algorithms

Finally, we define here the clustering algorithm used in this work to disambiguate entities globally.

2.8.1 Hierarchical Agglomerative Clustering (HAC)

The hierarchical agglomerative clustering model works by building clusters of items through sequences of merges. Initially, each item is placed in its own cluster, and these clusters are iteratively merged greedily based on a given similarity metric [8] [9], until either all items are in the same cluster, or until a stopping condition has been met. This process is elaborated on in Algorithm 1. If the algorithm proceeds until all items belong to the same cluster, meaningful clusters can then be created by re-partitioning the tree based on a given metric. There are several possible similarity metrics, including:

1. *Single-Link:* Similarity is defined as the the *maximum* pairwise similarity between items from two clusters.

2. *Complete-Link:* Similarity is defined as the *minimum* pairwise similarity between items from two clusters.

3. *Average Link:* Similarity is defined as the average pairwise similarity between items from two clusters.

   $$Sim(C_i, C_j) = \frac{1}{(N_i)(N_j)} \sum_{x \in C_i, y \in C_j} S(x, y) \tag{2.20}$$

   where N_i is the size of cluster i and N_j is the size of cluster j.

4. *Group Average:* Similarity is defined as the average pairwise similarity between items in a merged cluster.

   $$Sim(C_i, C_j) = \frac{1}{(N_i + N_j)(N_i + N_j - 1)} \sum_{x, y \in C_i \cup C_j, x \neq y} S(x, y) \tag{2.21}$$

   where N_i is the size of cluster i and N_j is the size of cluster j.

For example, to achieve an appropriate resolution, we might select a minimum group average similarity (maximum average cophenetic distance) as the threshold for partitioning each cluster.

The time complexity of the this algorithm is $O(n^3)$, and the space complexity is $O(n^2)$, where n is the number of items to be clustered.

Algorithm 1 Hierarchical Agglomerative Clustering Algorithm

```
procedure INITIALIZE
    // Assign each item to its own cluster.
    for item $I_i \in \{\text{items}\}$ do
        $C_i \leftarrow I_i$
procedure ITERATE
loop:
    // Greedily merge clusters.
    $S = \{\}$ // Keep track of cluster similarity scores.
    for cluster $C_i \in \{\text{clusters}\}$ do
        for cluster $C_j \in \{\text{clusters - } C_i\}$ do
            $S_{i,j} \leftarrow Sim(C_i, C_j)$
    $C_i, C_j \leftarrow max(S)$ // Select the two clusters with the highest similarity.
    if $Sim(C_i, C_j) \geq t$ then
        // If their similarity meets the required threshold, merge the two.
        $Merge(C_i, C_j)$
        goto loop
```

2.9 Similarity Measures

We present here two commonly used string similarity measures that we use in this work.

2.9.1 Levenshtein Distance

An appropriate metric for detecting spelling/typographical errors between two strings is the Levenshtein distance. The Levenshtein distance, also referred to as *edit distance*, represents the minimum number of character edits that is required to transform one string into another [17]. A character edit is an insertion, deletion, or substitution. For example, the Levenshtein distance between the strings "Pfizer Inc" and "Pfizer inc." is 2, because 1 character substitution, and 1 character deletion (or addition) is required.

More formally, the Levenshtein distance between two strings, a and b, is defined as:

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} max(i,j) & \text{if } \min(i,j) = 0, \\ min \begin{cases} \mathrm{lev}_{a,b}(i-1,j)+1 \\ \mathrm{lev}_{a,b}(i,j-1)+1 \\ \mathrm{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases} \tag{2.22}$$

where $1_{(a_i \neq b_j)} = 1$ when $a_i \neq b_j$ and 0 when $a_i = b_j$.

2.9.2 N-Grams

An n-gram is a sequence of n word tokens from a source sequence, which in our context is an entity mention string, tokenized based on whitespaces and punctuation. An n-gram with $n = 1$ is referred to as a unigram, $n = 2$ a bigram, and $n = 3$ a trigram. For example, consider the following two entity mentions:

1. E_1 = “Glaxo Smith Kline”
2. E_2 = “Glaxo Smith Kline Pharmaceuticals”

The n-grams for E_1 would be:

- unigrams: $\{Glaxo, Smith, Kline\}$
- bigrams: $\{\{Glaxo, Smith\}, \{Smith, Kline\}\}$
- trigrams: $\{\{Glaxo, Smith, Kline\}\}$

Similarly, the n-grams for E_2 would be:

- unigrams: $\{Glaxo, Smith, Kline, Pharmaceuticals\}$
- bigrams: $\{\{Glaxo, Smith\}, \{Smith, Kline\}, \{Kline, Pharmaceuticals\}\}$

- trigrams: $\{\{Glaxo, Smith, Kline\}, \{Smith, Kline, Pharmaceuticals\}\}$

To construct a meaningful feature from n-grams, we calculate their Jaccard index. The Jaccard index for the unigrams of this example would be as follows:

$$\text{Jaccard}(E_1, E_2) = \frac{|\{Glaxo, Smith, Kline\}|}{|\{Glaxo, Smith, Kline, Pharmaceuticals\}|} = \frac{3}{4} = 0.75$$

CHAPTER 3
Problem Definition, Methods, and Data

In this work, we begin with a conflict of interest dataset, sampled from the top ten medical journals by impact factor. The dataset consists of a set of biomedical articles and their authors, as well as labelled entity mentions and relations from the conflict of interest statements. Furthermore, each article is double-coded. This means that for each article, there are two independently labelled sets of the entity mentions and relations. Finally, along with each labelled entity mention, we also have a labelled entity type. Below is an example conflict of interest statement:

> *AUTHOR_1 reports having received research grant support and lecture fees from Bristol-Myers Squibb (BMS) and Merck; AUTHOR_2 grant support from Novartis; AUTHOR_3 grant support and lecture fees from Novartis, as well as lecture fees from BMS and Pfizer; and AUTHOR_4 consulting fees from Novartis and lecture fees from Pfizer and Novartis. Brigham and Women's Hospital has been awarded patents regarding the use of inhibition of the renin-angiotensin system in selected survivors of myocardial infarction; AUTHOR_1 and AUTHOR_3 are among the coinventors.*

Our dataset thus contains the labelled relations for this disclosure statement expressed in Table 3–1.

Table 3–1: Sample author-entity relations.

Relationship Type	**E-PER (Author)**	**E-ORG (Entity)**
received_research_grant_directly	AUTHOR_1	Bristol-Myers Squibb
received_research_grant_directly	AUTHOR_1	Merck
lecture_fees	AUTHOR_1	Bristol-Myers Squibb
lecture_fees	AUTHOR_1	Merck
patent	AUTHOR_1	Brigham and Women's Hospital
received_research_grant_directly	AUTHOR_2	Novartis
received_research_grant_directly	AUTHOR_3	Novartis
lecture_fees	AUTHOR_3	Novartis
lecture_fees	AUTHOR_3	BMS
lecture_fees	AUTHOR_3	Pfizer
patent	AUTHOR_3	Brigham and Women's Hospital
consulting_fees	AUTHOR_4	Novartis
lecture_fees	AUTHOR_4	Pfizer
lecture_fees	AUTHOR_4	Novartis

Since this dataset is double coded, we must handle (dis)agreement between the codings. It is possible, for example, that one annotator coded *lecture_fees(AUTHOR_3, BMS)*, whereas the other annotator coded *lecture_fees(AUTHOR_3, Bristol-Myers Squibb).* Given the text of the article, both of these candidate strings are reasonable, and we must identify that "Bristol-Myers Squibb" and "BMS" represent the same real-world entity. There are several sources for discrepancies between entity mentions across the two codings, as well as other discrepancy types, such as differing relationship types. However, in this context, we are only concerned with handling discrepancies between entity mentions.

Now, consider an excerpt from another conflict of interest statement:

> *AUTHOR_5 has received research funding from the MDA, NIH/NINDS, National ALS Association, Novartis Pharmaceuticals, and Alexion Pharmaceuticals.*

Given this statement, we generate another set of author-entity relations, and observe that AUTHOR_5 has received funding from "Novartis Pharmaceuticals". If we were to naively generate a network given these two entities and their related authors, we would produce the network depicted in Figure 3–1.

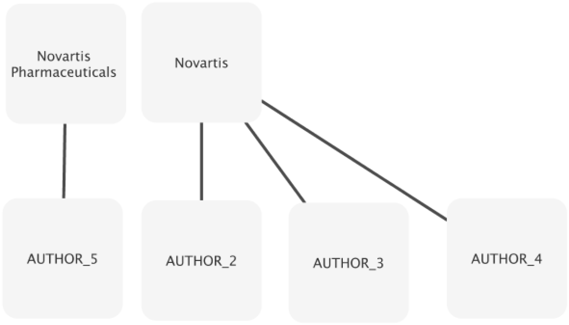

Figure 3–1: Sample entity-author network prior to disambiguation.

Any insights from analyzing this graph would then be skewed, as the network structure does not accurately represent the real-world engagement network, where all five authors have a relationship with a common entity. In this case, the "Novartis" and "Novartis Pharmaceuticals" nodes should be collapsed into a single node.

Thus the process of named entity disambiguation for this dataset is twofold:

1. **Local Disambiguation**: identifying coreferent entity mentions, from within the same article, and conditioned on the associated author and relationship type. This yields coreferent entity mention dyads.

2. **Global Disambiguation**: identifying coreferent entity string dyads from the codings of all articles, to form clusters of coreferent entity mentions.

Note that we have a unique identifier associated and linked with each author. As a result, no disambiguation is required for authors.

3.1 Entity Disambiguation Objectives

We define here several objectives for disambiguating entities, that apply during both folds of the disambiguation process. These rules have been written with a "conservative" policy on linking entities, so as not to overstate claims or invalidate any findings derived from the resulting dataset. That is to say that *not* linking a coreferent entity dyad is preferable to linking a non-coreferent entity dyad. This approach favours leaving the data "as is" (at a higher resolution), rather than intervening in ambiguous cases.

We define here 6 generalized sources of discrepancies between entity mentions, and how our model should handle these scenarios.

1. **Spelling/Typographical Error**: Match an entity whose name is clearly a typographical error to the correct entity. These discrepancies include cut-off characters, leading or trailing whitespaces, or other errors in spelling.
   - Eg. "Novartis" vs. "Novarti"
2. **Acronyms & Initialisms**: These discrepancies occur due to initialisms or acronyms for a given entity. Match an entity that is (or contains) an initialism, or is (or contains) a shortened form of a word, with the full entity name.
   - Eg. "National Institutes of Health" vs. "NIH"

3. **Abbreviations**: These discrepancies occur due to shortened forms of certain terms.
   - Eg. "Pfizer Inc" vs. "Pfizer Incorporated"
4. **Entity resolution**: These discrepancies occur due to the specifity of an entity or one of its subsidiaries. The rationale in this scenario is that the declaration of particular entity resolution is up to the discretion of each individual author. Some may elect to use coarse-grained names, such as "NIH", while others might prefer more fine-grained resolution that includes a subsidiary institute, department, etc. We cannot determine whether an author naming "NIH" actually has a relationship directly with "NIH", or with a subsidiary, and simply elected not to give more fine-grained information. Therefore, we default to the parent entity in all cases.
   - Eg. "Brigham and Women's Hospital" vs. "Brigham and Women's Hospital, Division of Cardiovascular Medicine"
5. **Geographic differentiation**: Match a geographically-specified entity to its parent entity (geographically unspecified). As for entity resolution above, geographic specificity is up to the discretion of each author.
   - Eg. "Pfizer Ltd" vs. "Pfizer GmbH, Karlsruhe, Germany"
6. **Partnerships**: If one entity name contains two or more already recognized entity names, do not match to any of the names.

3.2 Local Entity Disambiguation

For local entity disambiguation, we are concerned with disambiguating entity mentions from within the same conflict of interest text. Discrepancies are inevitable, as a result of the hand-coded nature of the dataset. However, we are able to exploit the local nature of the data. Because the entity mentions are extracted from the same source text, we can expect the entity string discrepancies to be due to human error, as opposed to fundamental differences in the way the entity mentions are reported.

Given the above categorical sources of error, we implement a binary entity linker to classify dyads of entity mentions. These entity mentions come from the double coded article data, and are conditioned on the relationship type and author, as well as the entity type. This means that only entity mentions with a common author and relationship type are considered for linking, and thus not all possible combinations of entity dyads are considered for matches. We consider only entity dyads where a potential match is signaled by a common relationship with an author.

3.3 Local Entity Linker

3.3.1 Preprocessing

We first leverage several string preprocessing techniques to normalize the entity mention strings, to enhance the performance of the model. To account for abbreviations in the entity strings, a dictionary of common abbreviations was used to expand abbreviated words and certain punctuation. This normalizes terminology, such that tokens with an equivalent meaning are treated as such. For example, "Inc" is replaced with "Incorporated", and ampersands are expanded to the word "and". By replacing tokens equivalent in meaning with a common token, it increases the overlap of n-grams and other distance metrics, as the tokens become equivalent. We also strip all leading and trailing whitespaces from entity mentions. Finally, the mention strings are set to lowercase.

3.3.2 Model Features

We present here the model features used in the entity linker, constructed with the entity disambiguation objectives in mind.

Levenshtein Distance

We use the Levenshtein distance, as presented in Chapter 2, as a feature in our model, for resolving minute typographical or spelling errors that result in entity mention discrepancies.

N-Gram Overlap

The Jaccard index of each of the unigrams, bigrams, and trigrams are used as features in the model, to capture overlap between any two given entity strings. Note that we need not include the Jaccard distance as a feature, as it is complimentary to the Jaccard index.

Acronyms/Initialisms

Identifying initialisms in entity strings is important, as these can result in a low n-gram intersection, and a fairly large edit distance between two entity mentions, rendering the features above inadequate on their own. For example, "NIHR" should match with "National Institutes of Health Research", however there is no n-gram overlap, and the Levenshtein distance is very large, at 34.

To identify acronyms and initialisms in the entity mentions, we identify all fully capitalized tokens in one entity string, and search for a sequence of entity tokens in the other entity string where the starting characters match. For example, for the two entity mentions "**CIHR**" and "**C**anadian **I**nstitutes of **H**ealth **R**esearch", we identify **CIHR** as an initialism because it is fully capitalized. We then search the second entity string for a sequence of tokens beginning with the the capitalized characters "C-I-H-R", ignoring stopwords. The full algorithm is presented in Algorithm 2, and is used symmetrically on each entity dyad.

We use a binary feature in our model to indicate if there is an acronym match for either entity string, taking a value of true if an acronym was identified, and false otherwise.

Algorithm 2 Algorithm for Detecting Acronyms/Initialisms

```
procedure FindAcronymsAndInitialisms
    // Tokenize each entity, removing punctuation and stop words.
    entity_1_tokens ← tokenize(entity_1, remove_stopwords=True)
    entity_2_tokens ← tokenize(entity_2, remove_stopwords=True)
    for e1_token in entity_1_tokens do
        if e1_token is uppercase and len(e1_token) >= 3 then
            for e2_token in entity_2_tokens do
                if e1_token[0] == e2_token[0] then
                    // Record this offset as a potential initialism starting point
                    offsets ← index(e2_token)
            // Iterate over each potential initialism starting point
            for offset in offsets do
                match ← True // Assume initialism match
                // Iterate over each character in the token for entity 1
                for entity_1_char, entity_1_char_index in entity1_token do
                    // Retrieve the corresponding token in entity 2
                    e2_token ← entity_2_tokens[entity_1_char_index + offset]
                    // Check for a character match
                    if entity_1_char[0] != e2_token[0] then
                        The sequence has been broken; no initialism for this offset.
                        match ← False
                // An initialism was encountered for the current offset.
                if match == True then
                    return True
    return False
```

Substrings

We also consider if one entity mention is a direct substring of the other, and use a binary feature to encode this value. To do so, we search for a sequence of tokens in the longer of the two entity mentions that perfectly matches the entire sequence of tokens in the shorter entity mention. A value of true if is assigned if a substring is identified, and false otherwise.

Adjectives of Countries and Nations

Finally, we consider the names of countries or nations contained in the entity mentions, in either noun or adjective form.

For example, take the two entity mentions, "Dutch Ministry of Health" and "Kenyan Ministry of Health". Each entity mention is similar across other features previously discussed, though they are not coreferent. This is evident as each entity mention contains an adjectival form of the country in which the entity resides.

To identify scenarios such as this, we search for the names of any country, as well as the adjectival forms of those countries, in each entity mention. If an adjective match is detected, the adjectival form is mapped back to the corresponding country. The countries identified in each entity mention are then compared, and the results are then encoded as a one-hot feature in the model. The three options are:

1. **Match**: Both entity mentions contain a country name or adjective, and both resolve to the same country.
2. **Non-Match**: Both entity mentions contain a country name or adjective, however both resolve to the differing countries.

3. **Ambiguity:** One of the mentions contains a country name or adjective, and the other entity mention does not.

3.3.3 Model Training

Given the above features for each candidate entity mention pair, we train each of the classification models presented in Section 2.7, using a grid search to find the optimal hyper-parameters and k-fold cross validation for evaluation. The results are presented in Section 4.1.

3.4 Global Entity Disambiguation

The local entity disambiguation process yields pairs of coreferent entity mentions. In cases where the entity linker paired non-identical entity mentions, we have an author or study relationship with a pair of entity mentions, which we refer to as an *entity dyad*. These entity dyads are produced locally in each article, though for the dataset to be of any use, they must be disambiguated against all other entity dyads from the articles, using a clustering algorithm.

For this global clustering problem, we employ the hierarchical agglomerative clustering model presented in Section 2.8, using group average similarity as the cluster scoring metric.

3.4.1 Entity Dyad Pairwise Similarity

Hierarchical agglomerative clustering requires a pairwise distance function between items (entity dyads) to form the clusters. We define this function using the entity linker presented previously in Section 3.3. Because each entity dyad has two entity mentions, there are 4 pairwise comparisons to be made between two entity dyads, as depicted in Figure 3–2. For each pair of entity dyads, we use the entity linker to determine these pairwise probabilities. For our dyad similarity function, we average the probabilities of the positive class. This is defined mathematically as:

$$\mathrm{Sim}(D_1, D_2) = \sum_{e_1 \in D_1} \sum_{e_2 \in D_2} \frac{P(e_1, e_2)}{4} \tag{3.1}$$

where $P(e_1, e_2)$ is the probability from the entity linker that the two entity mentions are coreferent. We then use this similarity function to define the pairwise entity dyad distance as:

$$\mathrm{Dist}(D_1, D_2) = 1 - \mathrm{Sim}(D_1, D_2) \tag{3.2}$$

Both these functions have a range of [0, 1]. A similarity value of 1.0 entails that all four entities from the two entity dyads are a considered a perfect match by the entity linker. A value of 0 entails that all four entities are considered perfect non-matches.

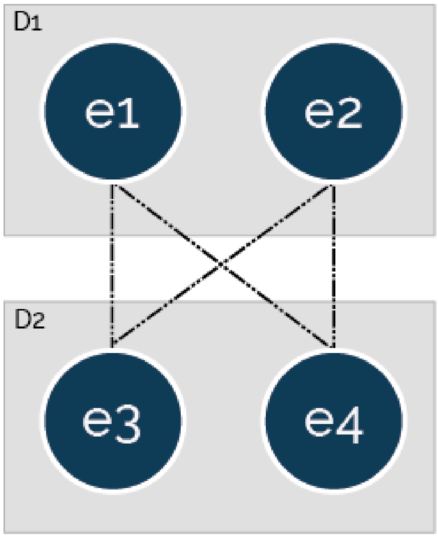

Figure 3–2: Entity dyad similarity comparisons.

3.4.2 Global Entity Clustering

As discussed, we use the hierarchical clustering method presented above to cluster the entity dyads, and use group average similarity to partition the clusters, experimenting with various threshold values t. We use the function in Equation 3.2 as the pairwise dyad distance function, as our implementation uses distance rather than similarity [14]. Finally, to improve both the quality of the clusters, as well as to improve the computational requirements of the algorithm, we leverage the labelled entity classifications discussed in Section 2.5.1, and block clustering of entity dyads with conflicting types. This yields a set of entity clusters for each entity type of corporation, academic, government, foundation, hospital, professional association/organization, and partnership. Partitioning the data in this way enables us to tune the clustering thresholds for each entity type, and observe differences among types.

CHAPTER 4
Evaluation and Results

4.1 Local Entity Disambiguation

To train and test the local entity linker, a set of coreferent entity mentions from across articles was manually coded. A subset of the 15,505 entity mentions was randomly sampled from the set of all entity mentions, and coded against each other manually by an expert coder, to identify pairs of coreferent entities. Coding was performed according to the disambiguation objectives described in Section 3.1. Due to the nature of this data, there is a large class imbalance between coreferent entity mention pairs and those that are not. The labelled dataset contains 6500 entity mention pairs, of which 1000 are coreferent and 5500 are not. The data are randomly under-sampled [16] to achieve a balanced dataset of 50% positive and 50% negative samples. This is a technique performed to increase the sensitivity of the classifier to the positive class, with the consequence that the local metrics may differ from the "real world" scenario. This is justified however, in that local entity disambiguation is an intermediate step in the global disambiguation process.

4.1.1 Test Results

Each classification model presented in Section 2.7 was trained on this data using the features from Section 3.3, using a grid search to find the optimal model parameters, and stratified k-fold for cross validation. We also experiment with adjusting the minimum confidence required to consider two entities as a match. In general, the convention is to predict the class with the maximum probability ($p > 0.50$ for binary classification). However, to increase model precision, we experiment with various values of p, using a discrete range of {0.50, 0.55, 0.6, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.9}. We select as our optimal model the model with the highest F_1 score across all configurations from the grid search. The results of each model are shown in the following sections.

Naïve Bayes

The optimal Naïve Bayes model, determined through a hyper-parameter grid search, used a Gaussian distribution for the data, out of Gaussian, multinomial, and Bernoulli. This model had a mean precision of 86.7%, mean recall of 96.3% and mean F_1 score of 91.1%.

Logistic Regression

The optimal logistic regression model, determined through a hyper-parameter grid search, utilized L2 regularization, with a regularization parameter, λ, of 1.25. This model had a mean precision of 96.4%, mean recall of 92.5% and mean F_1 score of 94.4%, the best of the models, and the model used to generate entity dyads for the global disambiguation process.

Decision Tree (CART)

The optimal decision tree, determined through a hyper-parameter grid search, utilized Gini impurity as the objective splitting function, with a minimum leaf node size of 1 and no maximum tree depth. This model had a mean precision of 92.5%, mean recall of 91.2% and mean F_1 score of 91.8%.

SVC

The optimal support vector classification model, determined through a hyperparameter grid search, utilized a polynomial kernel of degree 3, with a regularization parameter of 1.0. Kernel functions considered are linear, sigmoid, polynomial, and radial basis functions. This model had a mean precision of 94.3%, mean recall of 94.1% and mean F_1 score of 94.1%.

Table 4–1: Naïve Bayes local entity linking results.

Fold	**Accuracy**	**Precision**	**Recall**	$\mathbf{F_1}$
1	0.955	0.869	0.933	0.900
2	0.963	0.829	0.964	0.892
3	0.997	0.944	1.0	0.971
4	0.994	0.935	0.960	0.947
5	0.954	0.756	0.957	0.845
Mean	0.973	0.867	0.963	0.911
Std. Dev.	0.0188	0.0698	0.0215	0.0445

Table 4–2: Logistic Regression local entity linking results.

Fold	**Accuracy**	**Precision**	**Recall**	$\mathbf{F_1}$
1	0.953	0.923	0.851	0.886
2	0.980	0.989	0.883	0.933
3	1.000	1.000	1.000	1.000
4	0.997	0.986	0.960	0.973
5	0.981	0.921	0.932	0.926
Mean	0.982	0.964	0.925	0.944
Std. Dev.	0.0168	0.0344	0.0532	0.0395

Table 4–3: Decision tree local entity linking results.

Fold	**Accuracy**	**Precision**	**Recall**	$\mathbf{F_1}$
1	0.938	0.872	0.836	0.854
2	0.970	0.934	0.8722	0.902
3	0.995	0.943	0.971	0.957
4	0.996	0.986	0.947	0.966
5	0.976	0.888	0.932	0.910
Mean	0.975	0.925	0.912	0.918
Std. Dev.	0.0211	0.0407	0.0496	0.0405

Table 4–4: Support vector classifier local entity linking results.

Fold	**Accuracy**	**Precision**	**Recall**	$\mathbf{F_1}$
1	0.957	0.912	0.885	0.898
2	0.981	0.967	0.908	0.937
3	0.997	0.944	1.000	0.971
4	0.997	0.986	0.960	0.973
5	0.981	0.9059	0.9506	0.9277
Mean	0.982	0.943	0.941	0.941
Std. Dev.	0.0147	0.0310	0.0404	0.0282

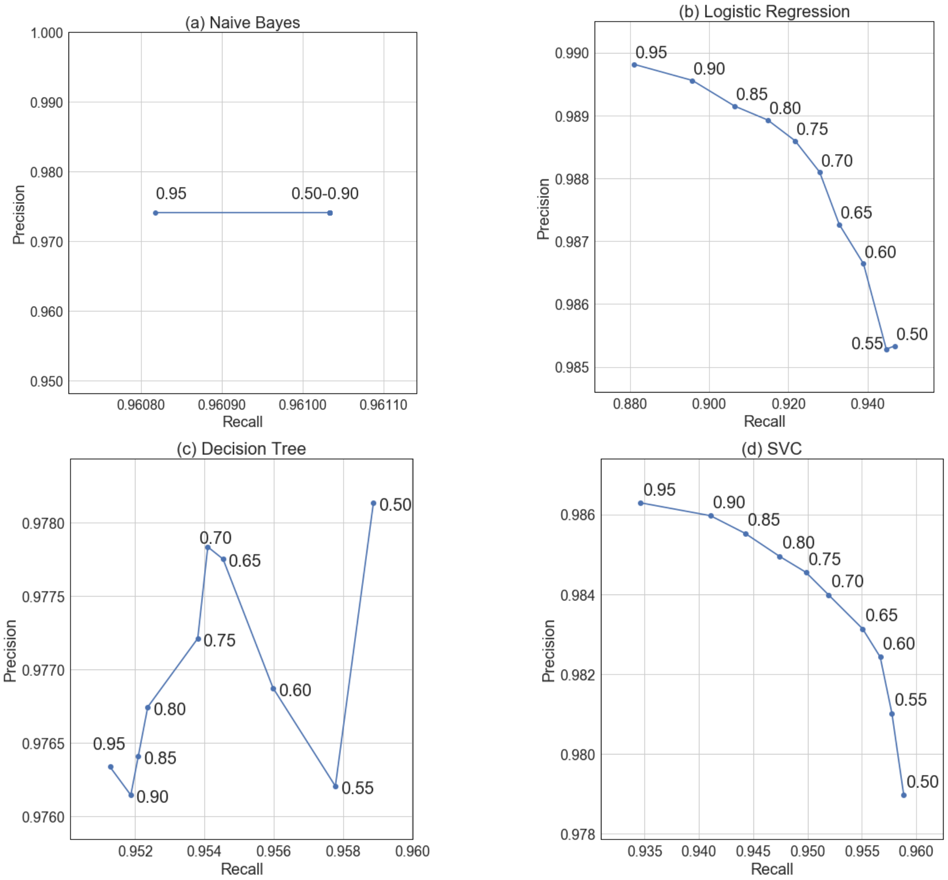

Figure 4–1: Precision-recall curves for Naïve Bayes, Logistic Regression, Decision Tree and SVC models, for varying probability thresholds.
The scales for each plot have been chosen to present the shape of each curve, due to the different regions where they reside.

4.1.2 Model Selection

The two top performing models are support vector classifiers and logistic regression, with naive Bayes and decision trees under-performing relative to these models. Between the SVC and logistic regression however, we see that while logistic regression has a marginally higher average F_1 score, this metric also has a marginally higher variance. For these reasons, the use of either model for this task can be justified.

However, we choose logistic regression due to its higher performance and simplicity over the support vector classifier.

4.1.3 Experimental Results

In practice, the local entity linker is used to identify coreferent entity mentions in the text, at an article level, as discussed in Chapter 3. Because these entity mentions are conditioned on relationships with authors, there is a higher chance that entity mention pairs are indeed coreferent, as they have also been conditioned on other criteria.

To gauge the performance of the entity linker in this operational context, we fully disambiguate all entities at a local level, conditioned on relationship type, using the optimal logistic regression model. We then randomly sample a subset of the resulting entity pairs for manual validation. This data gives us a practical measure of the number of false positives, though it does not provide a measure for false negatives, in the same way that the k-fold cross validation does. This is justified however, as false negatives are addressed by the cross validation process, and this metric is supplementary. From the data points, 150 unique entity dyads were randomly sampled and manually coded as described. Of these 150 unique entity dyads, conditioned also on an author relationship, 1 was a false positive. This is a 99.33% accuracy, excluding the intangible false negatives.

4.2 Global Entity Disambiguation

To perform the global entity disambiguation task, we use the optimal binary entity linker from the local disambiguation task, using the match probability in the dyad distance function of the hierarchical agglomerative clustering model. We use group average similarity and experiment with various values of t, such that clusters will only be merged if the resulting cluster has a group average similarity $\geq t$. To evaluate the resulting entity clusters, we use internal and external evaluation metrics.

4.2.1 External Evaluation

To quantify cluster quality externally, we manually construct "gold standard" clusters for a targeted set of known real-world entities of each classification type, by exhaustively examining the data and clustering entities manually. This gold standard standard consists of 96 entity clusters across the 7 entity types, including a total of 1395 entity mention dyads. To produce these clusters, an entity dyad was randomly sampled from the set of dyads for each entity type. Other mentions of this dyad were then searched for exhaustively and added to the cluster. Note that this set of entities is distinct from the local disambiguation training/testing set, to prevent overfitting and biased results. We then use this gold standard data to validate the clustering results, by aligning the generated clusters with the gold standard clusters, and evaluating the results. We use three metrics for this evaluation: Jaccard index of the predicted cluster and the gold-standard cluster, the inclusion error of the predicted cluster, and the exclusion error of the predicted cluster. Inclusion error measures the quantity of entity mentions that should not belong to a cluster, but are included erroneously. Similarly, exclusion error measures the quantity of entity mentions that should belong to a cluster, but are erroneously excluded. Figure 4–2 depicts the interaction of these evaluation metrics.

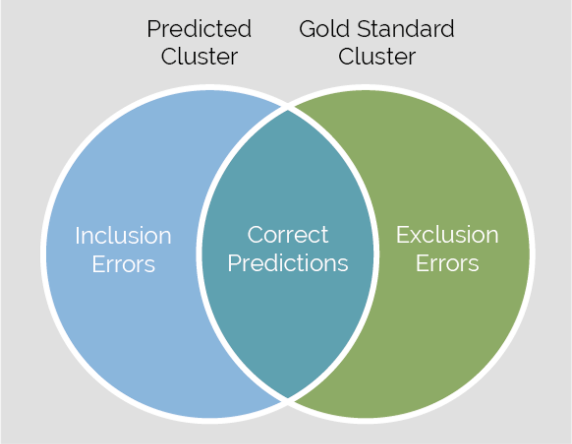

Figure 4–2: External cluster evaluation metrics.

In Figures 4–3 and 4–4, we plot the average of each of these metrics against the group average distance threshold used in the clustering algorithm, for each entity type. For the Jaccard index, we take a weighted average, such that larger clusters (from the gold standard) are weighted more heavily. Finally, we use evenly distributed distance threshold values in the range [0, 1], as this is the range of the distance function in Equation 3.2.

4.2.2 Internal Evaluation

To validate the clusters internally, we observe the average Silhouette coefficient across the generated clusters for each entity type, using the group average entity dyad distance function. This gives us a measure of the overall cohesion between clusters of each type. In Figure 4–5 we plot the average Silhouette coefficient against the distance threshold for each entity type.

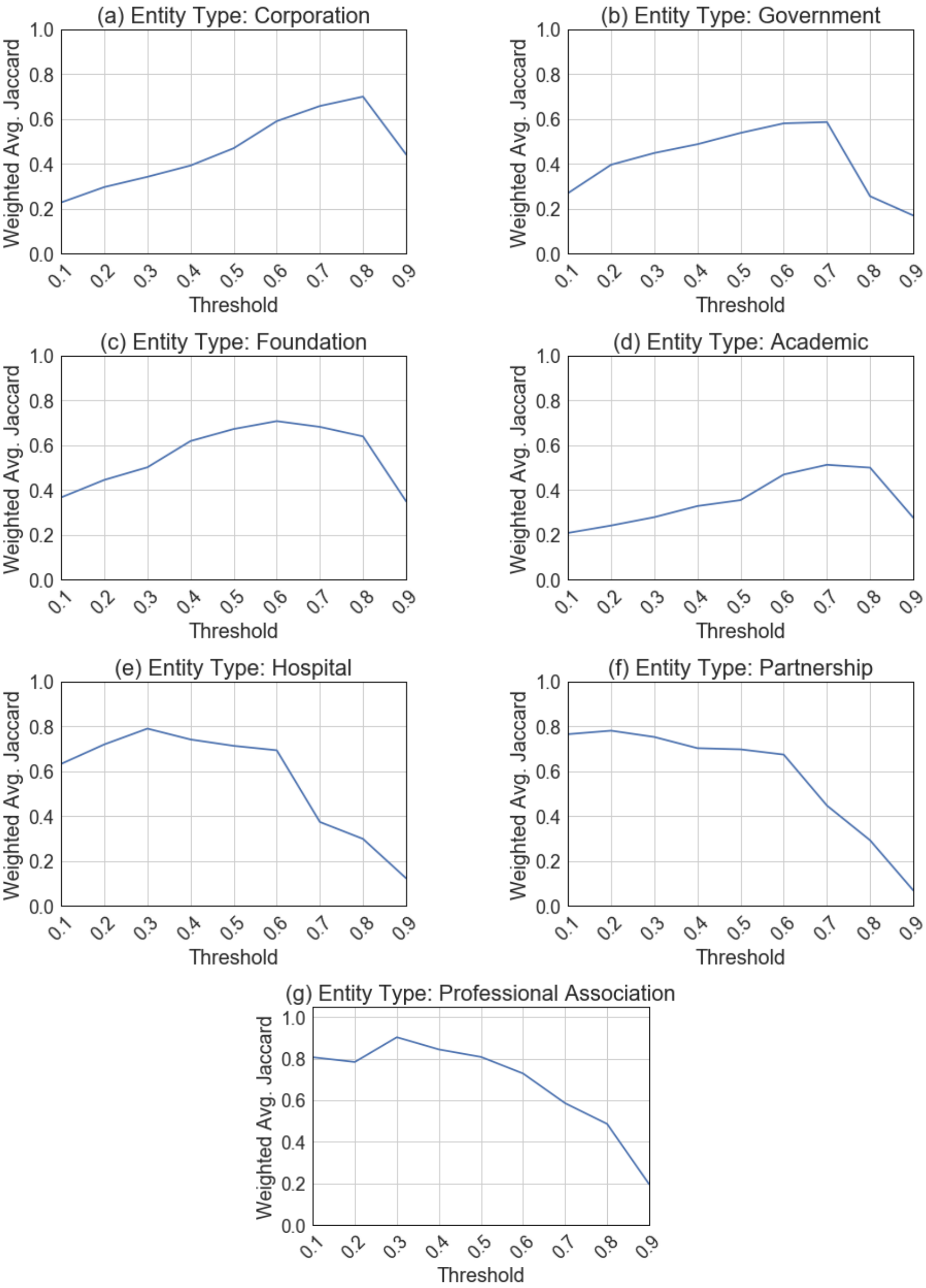

Figure 4–3: Average Jaccard coefficient for each entity type, weighted by the size of the gold-standard clusters, for varying distance thresholds.

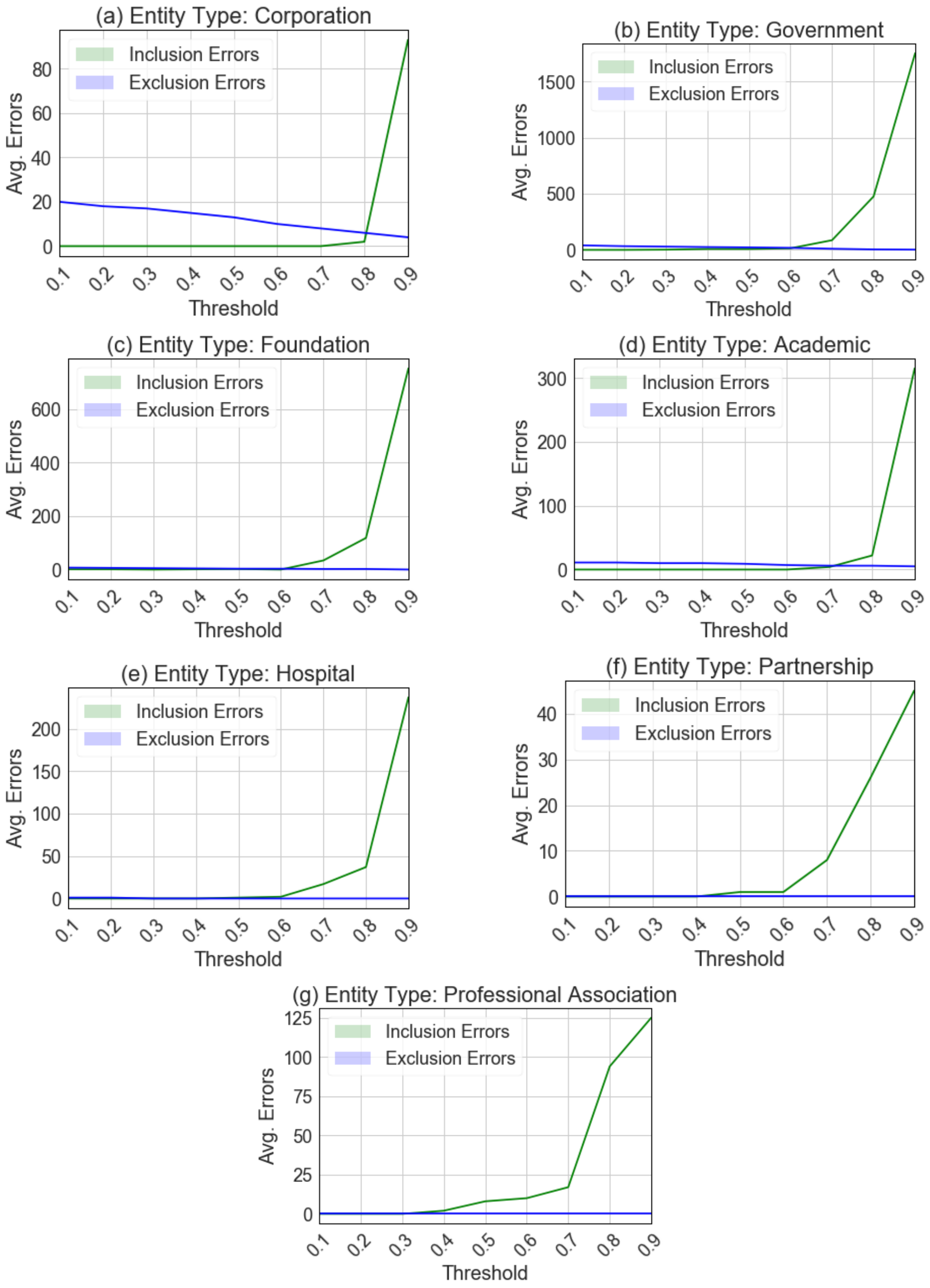

Figure 4–4: Absolute inclusion and exclusion errors for each entity type, for varying distance thresholds.

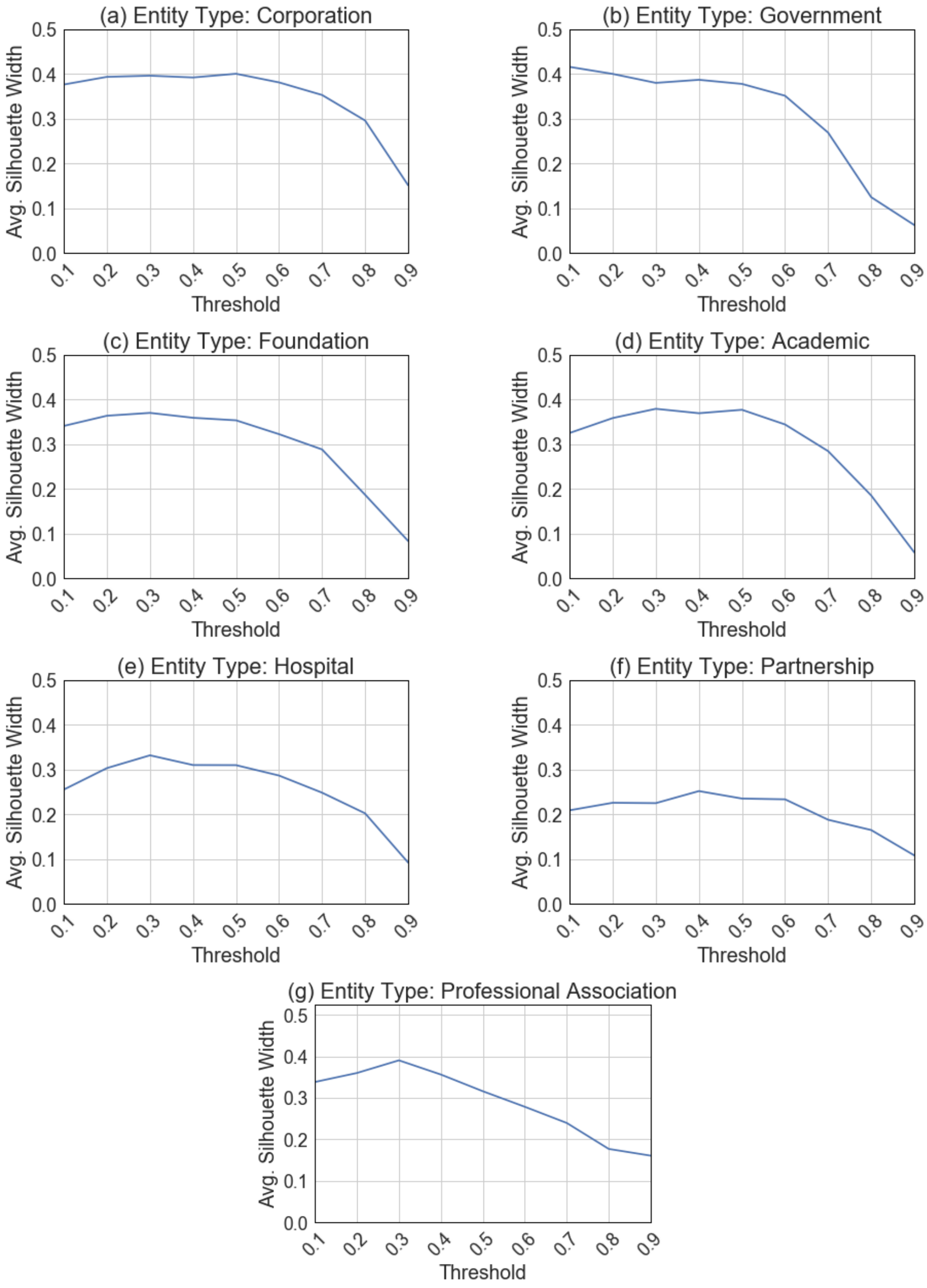

Figure 4–5: Average Silhouette width for each entity type, for varying distance thresholds.

CHAPTER 5
Discussion

5.1 Local Entity Disambiguation

Of the four entity linking models presented, we choose logistic regression with L2 regularization, and a regularization parameter of $\lambda = 1.25$. This model was selected as it has the highest average precision (96.4%) and F_1 (94.4%) of all models. Experimental testing at the local disambiguation level of this model, with a prediction threshold of $p = 0.50$ also yields very strong results, with 99.33% of the sampled linked entities being correctly classified. The strong performance of the entity linker is additionally bolstered by the nature of the task, where each article not only has a limited number of entity mentions, but also where the linking of these mentions is conditioned on entity-author relationships. These characteristics, when combined, yield exceptional results.

5.2 Global Entity Disambiguation

5.2.1 External Evaluation

At the global disambiguation level, we observe two trends for the external evaluation metrics of each entity type. For corporations, governments, foundations and academic institutions, we see the weighted average Jaccard score gradually increase to a peak, followed by a rapid decline at a threshold value close to 0.9. This also coincides with a rapid increase of inclusion errors at the same threshold value. This is intuitive, as the minimum group average distance threshold approaches the upper bound of the distance function (a threshold value of 1.0 would yield a single cluster containing all entities). For hospitals, partnerships and professional associations, we observe that this decrease in weighted average Jaccard scores, as well as the increase in inclusion errors, begins at lower threshold values, between 0.4 and 0.5. This trend can be explained intuitively via manual inspection of the clusters. We observe that for government organizations, hospitals and professional organizations, many entities have similar names, with overloaded naming schemes. For example, the entity mention "Medical Research Council" is more difficult to distinguish from "UK Medical Research Council" and "Kenyan Medical Research Council", than say the corporate entities "Pfizer" and "Novartis". These ambiguities arising from overloaded naming schemes pose an issue when disambiguating at a global level, and make it more difficult to cluster entities of these types than the others. This highlights a need for higher precision, at the expense of recall, for these entity types.

5.2.2 Internal Evaluation

The average silhouette width across clusters for each entity type provides a measure of cohesion across the clusters for each distance threshold. It is important to note that while internal cluster evaluation metrics provide insight into the structure and cohesion among clusters, this does not inherently validate the clustering results [18]. Hence, this metric is used in conjunction with the external evaluation metrics for cluster evaluation.

For the clusters of each entity type, we observe a relatively concave average silhouette width curve, approaching zero as the distance threshold approaches 1.0. An average silhouette width of under 0.20 is generally considered to be insufficient, and an average width of 0.50 or higher is generally considered reasonable [8], though silhouette width is more often used to tune clustering parameters. We observe that the threshold of 0.20 is adequately exceeded at the peaks of each curve, with the exception of the partnership entity type.

5.2.3 Distance Threshold Selection

To produce finalized clusters, we leverage both our internal and external evaluation metrics presented above. As stated previously, a conservative approach where exclusion errors are preferred over inclusion errors is optimal. However, the optimal threshold for inclusion errors across all entities is 0.1, and the optimal threshold for exclusion errors is 0.9, as these are the upper and lower bounds. Similarly, the weighted Jaccard index and average silhouette metrics have competing optimal thresholds, as shown in Table 5–1.

To select the optimal clustering for each entity type, we select a distance threshold balanced between the optimal weighted average Jaccard score and average silhouette width, weighting the weighted Jaccard score more heavily. The resulting thresholds are presented in Table 5–2.

We use these threshold values to cluster the data, and assign as the canonical name of each cluster the entity mention that appears in the most relationships in the data. Example clusters for each entity type can be found in Appendix A.

Table 5–1: Optimal thresholds for the weighted Jaccard and silhouette evaluation metrics.

Entity Type	**Weighted Jaccard**	**Silhouette**
Corporation	0.8	0.5
Government	0.7	0.1
Foundation	0.6	0.3
Academic	0.7	0.3
Hospital	0.3	0.3
Partnership	0.2	0.4
Professional Association	0.3	0.3

Table 5–2: Distance thresholds for each entity classification type.

Entity Type	**Threshold**
Corporation	0.7
Government	0.5
Foundation	0.5
Academic	0.6
Hospital	0.3
Partnership	0.3
Professional Association	0.3

CHAPTER 6
Application: Entity Influence Efficiency

In this chapter we present the results of an analysis that has motivated this research project: entity influence, and the identification of macro-scale strategies of engagement that entities use for engaging with authors. We hypothesize that entities, especially corporations, target "high-influence" authors for engagement, as they want to bias high-influence work in their favour.

6.1 Quantifying Influence

The influence of a researcher can be measured using various metrics, though a very commonly used metric is the h-index of the author [11]. The h-index of a given researcher is defined as their number of published papers, h, each with a number of citations $\geq h$. We use these indices for each author to gauge the potential influence of entities in our sample.

To do so, we propose an *influence efficiency* score for each entity in our sample. The influence efficiency of a given entity is the sum of the h-indices of each author that it engages with, divided by the maximum possible sum of h-indices for an entity with the same number of engagements, or the optimal score, for that year. It is a measure of how influential the engaged authors are for an entity, relative to other authors in the sample. It is important to emphasize that the h-score of an author changes over time, as they publish more papers and are cited more. Due to this, for each year that the author is observed in the sample, their h-score in that specific year is used. This value represents their influence at the time the engagement is reported, which is an approximation of when the engagement occurred. Mathematically, we define influence efficiency as:

$$\mathrm{IE}(E) = \frac{\sum_{a \in \mathrm{eng}(E)} h(a)}{\sum_{a \in \mathrm{eng}^*(E)} h(a)} \tag{6.1}$$

where $\mathrm{eng}(E)$ is the set of engaged authors for entity E, and $\mathrm{eng}^*(E)$ is the optimal set of engaged authors for entity E (the set of authors with the highest influence).

To perform this analysis, we scrape temporal h-index data for each author in our sample, using their unique SCOPUS identifiers. We then build a network of author nodes and entity nodes, using the disambiguated entity clusters from the previous chapter, and their coded relations to form edges between the nodes.

6.2 Expected Influence Efficiency

Using Equation 6.1 we can calculate the *actual* influence efficiency score of each given entity. However, to determine if these influence scores are significant, we must compare them against their *expected* values. The expected value is the score obtained by each entity if they were to randomly engage with researchers, with no preference for influence.

We implement this using a Markov Chain Monte Carlo (MCMC) algorithm to randomize the author-entity relationship network. Given the graph of entities and researchers, we generate 1000 random graphs using a degree preserved shuffle, and a semi-degree preserved shuffle, and calculate the mean influence efficiency scores and standard deviations for each entity. Given this, we calculate the single-tail p-value of the influence for each entity, with the hypothesis that entities actively target high-influence authors for engagements. These entity scores can then be leveraged to determine trends in the engagement patterns across entity types, using Fisher's combined probability test [3].

6.2.1 Random Graph Model

To ensure each generated graph is sufficiently randomized, we need to put a minimum bound on the number of random edge rewirings, N_R, that occur. For this, we use the following bound:

$$\mathrm{N}_R = E \times \ln \frac{1}{\epsilon}, \tag{6.2}$$

where E is the number of edges, and $\epsilon \leq 10^{-6}$ [25].

6.2.2 Degree Preserved Shuffle

The first random graph model considered is a degree preserved shuffle. This approach assumes that entities allocate a fixed amount of resources for engaging with authors, and that the number of authors they engage with remains constant. It also assumes that the degree of each researcher remains constant. This is perhaps not an accurate representation of how real-world engagements occur, as it is common for researchers to change or lose their engagements. However, it is an appropriate baseline.

6.2.3 Semi-Degree Preserved Shuffle

The second model considered is a modification of the degree preserved shuffle discussed above. In this model, the degree of each entity remains fixed, as they are assumed to have a fixed amount of resources for engaging with researchers. However, the number of entities that a researcher engages with, or the degree distribution of researcher nodes, is not preserved. This assumes that researchers are not inherently guaranteed a fixed number of engagements, and can gain or lose engagements to other researchers. This model is equivalent to randomly sampling n researchers for each entity in the sample, where n is the fixed number of engagements for that entity in that year, and treating those as the entity engagements.

6.3 Results

We first plot the standard deviations of the influence efficiency scores for each random graph model. We observe that the standard deviation plateaus as the number of engagements increases. This is intuitive, as an entity with few engagements has more combinations of authors to engage with, whereas an entity with many engagements has fewer combinations of potential authors. In Figures 6–1 and 6–2, we plot both the standard deviation of the influence efficiency scores, as well as the standard deviation of the influence efficiency scores normalized by the optimal influence score for that number of engagements in the given year. They are presented on both linear and log-linear plots. In Figure 6–3 we plot the optimal influence score that an entity can obtain for each number of engagements.

Due to the high variance of the influence scores for entities with a low number of engagements, we introduce a threshold such that we only consider entities with enough engagements that the variance is acceptable. This threshold is chosen to be at the shoulder of the curve in Figure 6–1, at a number of engagements ≥ 16. This threshold is plotted in red.

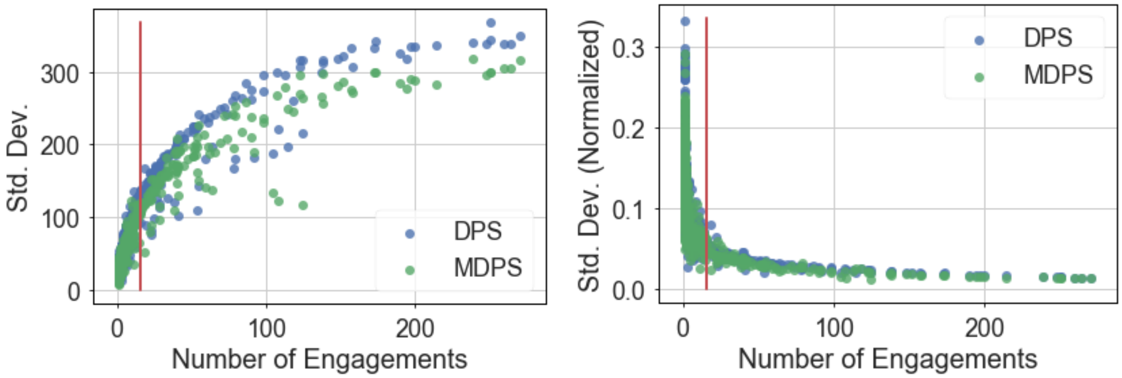

Figure 6–1: Influence efficiency score standard deviations.

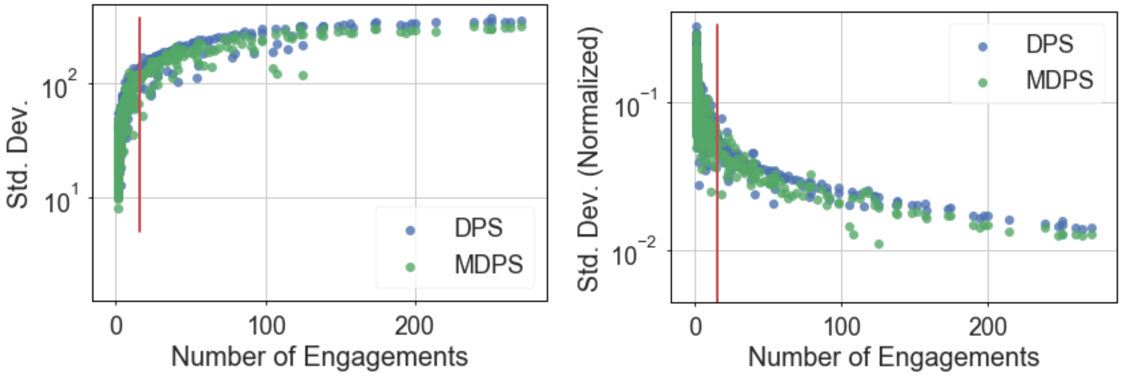

Figure 6–2: Influence efficiency score standard deviations on a log-linear plot.

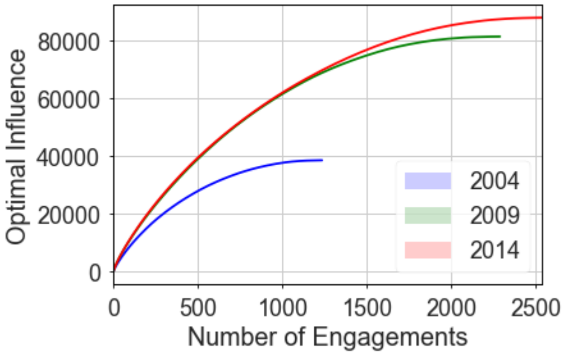

Figure 6–3: Optimal influence score.

Similarly, in Figure 6–4 we plot the p-values for each entity against the number of engagements. On the left we plot results for all entities, and the right for all entities where the number of engagements ≥ 16. Removing these entities prevents tests for academic institutions, hospitals, parternships and professional associations, as no entities of these entity types have more than 16 engagements in a given year. However, it serves to reduce the potential for spurious values, and increases our confidence in the results.

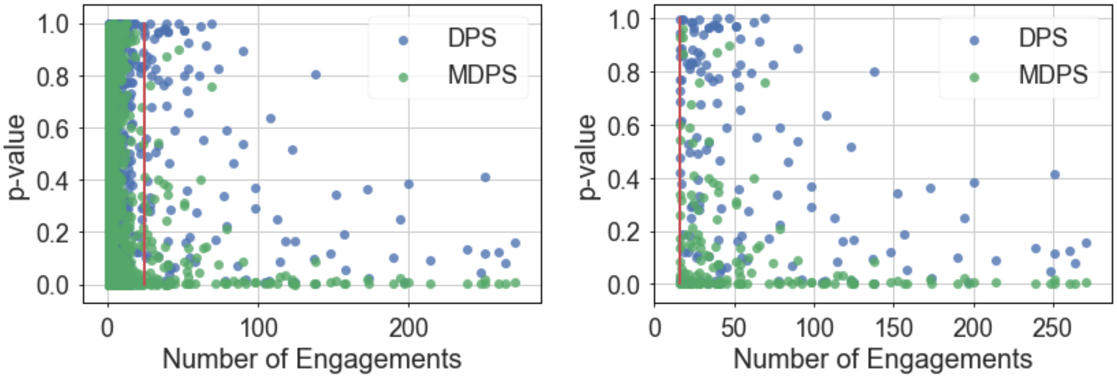

Figure 6–4: Individual entity influence efficiency significance.

Finally, we use Fisher's method of combining independent tests with a common hypothesis, to test our hypothesis that entities target high-influence authors. Fisher's method is defined as:

$$\chi^2 = \sum_{i=1}^{k} (-2 \log_e p_i) \tag{6.3}$$

where p_i is the p-value for entity i.

The results are reported in Tables 6–1 and 6–2, with significant results ($p \leq 0.05$) in bold.

Table 6–1: Influence efficiency degree preserved shuffle results across all years.

Entity Type	**2004**		**2009**		**2014**	
	Statistic	***p*-value**	**Statistic**	***p*-value**	**Statistic**	***p*-value**
Corporation	86.739	**0.00051983**	126.63	**0.0097287**	220.28	**2.3984e-10**
Government	1.1615	0.97874	5.2935	0.94745	3.9761	0.99999
Foundation	1.0681	0.8993	0.27429	0.99141	0.93304	0.98802

Table 6–2: Influence efficiency semi-degree preserved shuffle results.

Entity Type	**2004**		**2009**		**2014**	
	Statistic	***p*-value**	**Statistic**	***p*-value**	**Statistic**	***p*-value**
Corporation	177.82	**8.4789e-17**	463.64	**5.8644e-51**	566.07	**1.0429e-64**
Government	9.3663	0.154	28.409	**0.0048175**	39.594	**0.012056**
Foundation	3.9177	0.41726	4.3136	0.36522	10.677	0.098871

6.4 Discussion

The null hypothesis, that entities do not target influential authors for engagements, using their h-index as our proxy for influence, can be reasonably rejected for corporate entities across all observed years, and for both null network models. This is important, as it indicates that corporate entities are engaged with highly influential authors in their respective fields. However, it is also important not to construe this as proof of conflict of interest, as engagements do not inherently represent conflict of interest. It is intuitive for corporate entities to engage with highly influential authors, as they would expect quality results on their investments. Additionally, in the semi-degree preserved shuffle, the p-values decrease temporally, suggesting that corporate entities are gradually targeting their engagements more efficiently.

These results contrast the findings for government bodies and foundations. For these entity types, the null hypothesis can not be rejected for either the degree preserved shuffle or the semi-degree preserved shuffle, across the sampled years, with the exception of government entities for 2009 and 2014. This is also intuitive, as government bodies would be expected to support researchers more uniformly, and "share the wealth". Again, we see a temporal decrease in p-values for the semi-degree preserved shuffle for foundations, though not for government.

CHAPTER 7
Conclusions

7.1 Entity Disambiguation

7.1.1 Entity Classifications

First, we extend the LDC “Organization” entity classification type to better suit applications in a biomedical context. We then present how these entity types in our data can be leveraged to improve the entity disambiguation process, by enabling threshold optimization for each type. Finally, we are able to show how these classification types can inform industry engagement with authors in biomedical research.

7.1.2 Local Entity Disambiguation

We present a feature set for our entity linker that performs extremely well on the local disambiguation task, across all considered models. This indicates that the model features well capture the lexical and syntactic structures of named entity mentions in biomedical literature, enabling us to disambiguate them with confidence. Additionally, we show that this performance translates well in our experimental outcomes, where a set of linked entities was randomly sampled and manually coded, yielding strong results. We also observe that these experimental results are strengthened by the nature of the task, where entity linking can be conditioned on known relations between the entities and researchers.

7.1.3 Global Entity Disambiguation

We show that the entity linker output probabilities can be used effectively as a similarity function to disambiguate entities, using hierarchical agglomerative clustering. We implement this by averaging the pairwise probabilities of each entity across two entity dyads, and subtracting this from 1 to give us a distance in the range [0, 1]. We show that this model performs well across all entity classifications, though to varying degrees. We find that overloaded naming schemes can present issues for disambiguating entities, especially for government organizations, hospitals, and professional associations, though that this can be offset through threshold parameter tuning.

7.2 Applications

Finally, we present an analysis of entity engagement patterns in the field of conflict of interest, by defining an influence efficiency score, and computing it on aggregate for each entity type. This analysis presents evidence that corporate entities engage with higher influence researchers than can be expected randomly, and that government organizations and foundations do not to the same extent. This highlights a need for further investigation into the role these engagements take with researchers and studies, examining the presence of conflict of interest.

7.3 Conclusion

Conflict of interest remains a critical area of interest in our analysis and understanding of biomedical literature. This thesis presents methods for extracting information from research articles for the task of modelling conflict of interest, and presents methods to resolve the various difficulties that are inherent to the task, such as entity disambiguation and entity typing. Furthermore, we have produced a large-scale snapshot of entity-researcher engagements, that can be used in future efforts to study conflict of interest.

7.4 Future Work

Future work in named entity disambiguation in biomedical literature should consider different cluster distance metrics, such as single link or complete link, enabling tuning of this metric among entity types, and to target higher or lower resolution entity groupings, given a certain objective. This will enable fine-tuned targeted analyses of specific entities in the sample, and enable case-studies which in turn make the analyses more accessible. Work in conflict of interest should also present descriptive statistics on the presence of declared conflicts in biomedical research on aggregate, and investigate entity engagement strategies in more depth, by leveraging the known relationship types between researchers and entities. This would enable the characterization of relationship types and traits that are at a high risk for influence, and inform our understanding of how conflict of interest manifests.

Appendices

APPENDIX A
Example Entity Clusters

Table A–1: Corporation - *Boehringer Ingelheim* - Jaccard: 100.0%

Gold Standard Cluster	**Predicted Cluster**
Boehringer Diagnostic	Boehringer Diagnostic
Boehringer/Ingelheim	Boehringer/Ingelheim
Boehringuer-Ingelheim	Boehringuer-Ingelheim
Boehringer-Ingelhiem	Boehringer-Ingelhiem
Boehringer Ingelheim-Pfizer	Boehringer Ingelheim-Pfizer
Boehringer Ingleheim	Boehringer Ingleheim
Boehringer Ingelheim Pharmaceutical	Boehringer Ingelheim Pharmaceutical
Boehringer Ingelheim Korea	Boehringer Ingelheim Korea
Boehringer Ingelheim Pharma	Boehringer Ingelheim Pharma
Boehringer Ingelheim Stiftung	Boehringer Ingelheim Stiftung
Boehringer Mannheim Therapeutics	Boehringer Mannheim Therapeutics
Boehringer-Ingelheim	Boehringer-Ingelheim
Boehringer	Boehringer
Boehringer Diagnostic (Meylan)	Boehringer Diagnostic (Meylan)
Boehringer Ingelheim Pharmaceuticals, Inc	Boehringer Ingelheim Pharmaceuticals, Inc

Boehringer Ingelheim	Boehringer Ingelheim
Boehringer Ingelhiem	Boehringer Ingelhiem
Boehringer Ingelheim Espana	Boehringer Ingelheim Espana
Boehringer Ingelheim Pharmaceuticals Inc	Boehringer Ingelheim Pharmaceuticals Inc
Boehringer-Ingelheim Pharmaceutical	Boehringer-Ingelheim Pharmaceutical
Boehringer Ingelheim (Canada) Ltd.	Boehringer Ingelheim (Canada) Ltd.
Boehringer Ingelheim Pharmaceuticals	Boehringer Ingelheim Pharmaceuticals
Boehinger Ingelheim	Boehinger Ingelheim
Boeringher-Ingelheim	Boeringher-Ingelheim
Boerhinger-Ingelheim	Boerhinger-Ingelheim
Bohringer-Ingelheim	Bohringer-Ingelheim
Bohringer Ingelheim	Bohringer Ingelheim
Boeringer Ingelheim	Boeringer Ingelheim

Table A–2: Government - *FSRQ* - Jaccard: 92.9%

Gold Standard Cluster	**Predicted Cluster**
Chercheur-Boursier career award from the Fonds de recherche du Quebec-Sante	Chercheur-Boursier career award from the Fonds de recherche du Quebec-Sante
Clinician-Investigator Award from the Fonds du Quebec pour la Recherche en Sante	Clinician-Investigator Award from the Fonds du Quebec pour la Recherche en Sante

Clinician-Scientist Salary Award from the Fonds de la Recherche du Quebec-Sante	Clinician-Scientist Salary Award from the Fonds de la Recherche du Quebec-Sante
FRSQ	FRSQ
Fond de la Recherche en Sante du Quebec	Fond de la Recherche en Sante du Quebec
Fonds de Recherche en	Fonds de Recherche en
Sante du Quebec (FRSQ)	Sante du Quebec (FRSQ)
Fonds de la Recherche du Quebec-Sante	Fonds de la Recherche du Quebec-Sante
Fonds de la recherche en sante du Quebec	Fonds de la recherche en sante du Quebec
Fonds de recherche du Quebec-Sante	Fonds de recherche du Quebec-Sante
Fonds de recherche en sante du Quebec	Fonds de recherche en sante du Quebec
Fonds du Quebec pour la Recherche en Sante	Fonds du Quebec pour la Recherche en Sante
Reseau en Sante Respiratoire du Fonds de la Recherche en Sante du Quebec	Reseau en Sante Respiratoire du Fonds de la Recherche en Sante du Quebec
senior clinical-research scholar from the Fonds de la recherche en sante du Quebec	senior clinical-research scholar from the Fonds de la recherche en sante du Quebec
FRSQ Junior 1 Scholar	

Table A–3: Foundation - *Wellcome Trust* - Jaccard: 85.0%

Gold Standard Cluster	**Predicted Cluster**
Burroughs Wellcome Fund	
Career Award at the Scientific Interface	

Career Award at the Scientific Interface from the Burroughs Wellcome Fund	
Clinician Scientist Award from the Wellcome Trust	Clinician Scientist Award from the Wellcome Trust
Collaborative Research Initiative Grant from the Wellcome Trust (Grant reference: 063663/Z/01/Z), UK	Collaborative Research Initiative Grant from the Wellcome Trust (Grant reference: 063663/Z/01/Z), UK
Genotyping Facilities at the Wellcome Trust Sanger Institute	Genotyping Facilities at the Wellcome Trust Sanger Institute
Joint International Infectious Disease Initiative of the Wellcome Trust	
KEMRI-Wellcome Trust	KEMRI-Wellcome Trust
KEMRI-Wellcome Trust Research Programme's Clinical Trials Unit	KEMRI-Wellcome Trust Research Programme's Clinical Trials Unit
Research Leave Fellowship from The Wellcome Trust, UK	Research Leave Fellowship from The Wellcome Trust, UK
Scientific Interface	
Senior Research Fellowship from the Wellcome Trust	Senior Research Fellowship from the Wellcome Trust
Wellcome Senior Clinical Research Fellow	
Wellcome Trus	Wellcome Trus
Wellcome Trust	Wellcome Trust

Wellcome Trust	Wellcome Trust
Wellcome Trust (UK)	Wellcome Trust (UK)
Wellcome Trust 4-year studentship	Wellcome Trust 4-year studentship
Wellcome Trust Advanced Fellowship	Wellcome Trust Advanced Fellowship
Wellcome Trust Advanced Training Fellow	Wellcome Trust Advanced Training Fellow
Wellcome Trust Advanced Training Fellowship	Wellcome Trust Advanced Training Fellowship
Wellcome Trust Career Development Fellow	
Wellcome Trust Centre for Mitochondrial Research	Wellcome Trust Centre for Mitochondrial Research
Wellcome Trust Clinical Research Fellow	Wellcome Trust Clinical Research Fellow
Wellcome Trust Clinical Research Leave Fellow	Wellcome Trust Clinical Research Leave Fellow
Wellcome Trust Clinician Scientist Fellow	Wellcome Trust Clinician Scientist Fellow
Wellcome Trust Collaborative Research Initiative Grant	Wellcome Trust Collaborative Research Initiative Grant
Wellcome Trust Functional Genomics Initiative in Cardiovascular Genetics	
Wellcome Trust Intermediate fellowship in Public Health and Tropical Medicine	Wellcome Trust Intermediate fellowship in Public Health and Tropical Medicine

Wellcome Trust Masters Training Fellowship	Wellcome Trust Masters Training Fellowship
Wellcome Trust Principal Fellow	Wellcome Trust Principal Fellow
Wellcome Trust Principal Research Fellow	Wellcome Trust Principal Research Fellow
Wellcome Trust Principal Research Fellowship in the Clinical Sciences	Wellcome Trust Principal Research Fellowship in the Clinical Sciences
Wellcome Trust Sanger Institute	Wellcome Trust Sanger Institute
Wellcome Trust Seeding Drug Discovery Committee	Wellcome Trust Seeding Drug Discovery Committee
Wellcome Trust Senior Clinical Fellow	Wellcome Trust Senior Clinical Fellow
Wellcome Trust Senior Fellow	Wellcome Trust Senior Fellow
Wellcome Trust Senior Fellow in Clinical Science	Wellcome Trust Senior Fellow in Clinical Science
Wellcome Trust Senior Fellowship in Basic Biomedical Science	Wellcome Trust Senior Fellowship in Basic Biomedical Science
Wellcome Trust Senior Investigator	Wellcome Trust Senior Investigator
Wellcome Trust Senior Investigators	Wellcome Trust Senior Investigators
Wellcome Trust Senior Research Fellow	Wellcome Trust Senior Research Fellow
Wellcome Trust Senior Research Fellowship in Clinical Science	Wellcome Trust Senior Research Fellowship in Clinical Science
Wellcome Trust Strategic Award	Wellcome Trust Strategic Award

Wellcome Trust Training Fellow in Clinical Tropical Medicine	Wellcome Trust Training Fellow in Clinical Tropical Medicine
Wellcome Trust UK	Wellcome Trust UK
Wellcome Trust United Kingdom	Wellcome Trust United Kingdom
Wellcome Trust as a Research Training Fellow	Wellcome Trust as a Research Training Fellow
Wellcome Trust career development fellow	
Wellcome Trust core award	Wellcome Trust core award
Wellcome Trust of Great Britain	Wellcome Trust of Great Britain
Wellcome Trust of Great Britain (Major Overseas Programme-Thailand Unit Core Grant)	Wellcome Trust of Great Britain (Major Overseas Programme-Thailand Unit Core Grant)
Wellcome Trust project	Wellcome Trust project
Wellcome Trust research fellow	Wellcome Trust research fellow
Wellcome Trust senior fellow	Wellcome Trust senior fellow
Wellcome Trust senior research fellowship in clinical science	Wellcome Trust senior research fellowship in clinical science
Wellcome Trust's Tropical Interest Group	Wellcome Trust's Tropical Interest Group
Wellcome Trust,	Wellcome Trust,
Wellcome Trust, UK	Wellcome Trust, UK
Wellcome Trust, United Kingdom	Wellcome Trust, United Kingdom

Table A–4: Academic - *Karolinska Institutet* - Jaccard: 80.0%

Gold Standard Cluster	**Predicted Cluster**
Board of Postgraduate Education at Karolinska Institutet	Board of Postgraduate Education at Karolinska Institutet
Board of Postgraduate Education at Karolinska Institutet (Clinical Scientist Training Program; Dnr. 3023/11-225)	Board of Postgraduate Education at Karolinska Institutet (Clinical Scientist Training Program; Dnr. 3023/11-225)
Board of Postgraduate Education at Karolinska Institutet (Clinical Scientist Training Program; Dnr. 3023/11-225)	Board of Postgraduate Education at Karolinska Institutet (Clinical Scientist Training Program; Dnr. 3023/11-225)
Board of Research at Karolinska Institutet	Board of Research at Karolinska Institutet
Board of Research at Karolinska Institutet (Distinguished Professor Award; Dnr. 2368/10-221)	Board of Research at Karolinska Institutet (Distinguished Professor Award; Dnr. 2368/10-221)
Centre for Gender Medicine at Karolinska Insitutet	
Centre for Gender Medicine at Karolinska Institutet	Centre for Gender Medicine at Karolinska Institutet
Committee for Health and Caring Sciences and Strategic Research Program in Care Sciences at Karolinska Institutet	Committee for Health and Caring Sciences and Strategic Research Program in Care Sciences at Karolinska Institutet

Committee for Health and Caring Sciences and Strategic Research Program in Care Sciences at Karolinska Institutet (CT)	Committee for Health and Caring Sciences and Strategic Research Program in Care Sciences at Karolinska Institutet (CT)
Department of Medical Epidemiology and Biostatistics, Karolinska Institutet	Department of Medical Epidemiology and Biostatistics, Karolinska Institutet
Department of Medicine, Karolinska Institutet	Department of Medicine, Karolinska Institutet
Distinguished Professor Award from the Karolinska Institutet	Distinguished Professor Award from the Karolinska Institutet
Karolinska Cardiorenal Theme Center	
Karolinska Insitutet	
Karolinska Institute	
Karolinska Institute, Stockholm, Sweden	
Karolinska Institutet	Karolinska Institutet
Karolinska Institutet Center for Gender-based Research	Karolinska Institutet Center for Gender-based Research
Karolinska Institutet Funds	Karolinska Institutet Funds
Strategic Research Program in Diabetes at Karolinska Institutet	Strategic Research Program in Diabetes at Karolinska Institutet
Strategic Research Program in Epidemiology at Karolinska Institutet	Strategic Research Program in Epidemiology at Karolinska Institutet

Strategic Research Program in Epidemiology at Karolinska Institutet (Young Scholar Award; Dnr. 7340/2012)	Strategic Research Program in Epidemiology at Karolinska Institutet (Young Scholar Award; Dnr. 7340/2012)
Strategic Research Programme in Neuroscience at Karolinska Institutet (Strat-Neuro)	Strategic Research Programme in Neuroscience at Karolinska Institutet (Strat-Neuro)
strategic research program in epidemiology at Karolinska Institutet	strategic research program in epidemiology at Karolinska Institutet
strategic research programme in epidemiology at Karolinska Institutet	strategic research programme in epidemiology at Karolinska Institutet

Table A–5: Partnership - *Canadian Cardiovascular Outcomes Research Team* - Jaccard: 60.0%

Gold Standard Cluster	**Predicted Cluster**
Canadian Cardiovascular Outcomes Research Team	Canadian Cardiovascular Outcomes Research Team
Canadian Cardiovascular Outcomes Research Team (CCORT [www.ccort.ca])	Canadian Cardiovascular Outcomes Research Team (CCORT [www.ccort.ca])
Canadian Cardiovascular Outcomes Research Team Masters student fellowship	Canadian Cardiovascular Outcomes Research Team Masters student fellowship
	Alliance for Canadian Health Outcomes Research in Diabetes (ACHORD)

	Primary Care Outcomes Research Fellowship

Table A–6: Hospital - *Massachusetts General Hospital* - Jaccard: 84.6%

Gold Standard Cluster	**Predicted Cluster**
Center for Biostatistics, Massachusetts General Hospital	Center for Biostatistics, Massachusetts General Hospital
Eliot B. Shoolman Fund of the Massachusetts General Hospital	
Eliot B. and Edith C. Shoolman fund of the Massachusetts General Hospital	
Institute for Health Policy, Massachusetts General Hospital	Institute for Health Policy, Massachusetts General Hospital
Massachusetts General Hospital (MGH)	Massachusetts General Hospital (MGH)
Massachusetts General Hospital Anticoagulation Management Services	Massachusetts General Hospital Anticoagulation Management Services
Massachusetts General Hospital Center for D-receptor Activation Research	Massachusetts General Hospital Center for D-receptor Activation Research
Massachusetts General Hospital General Clinical Research Center	Massachusetts General Hospital General Clinical Research Center
Massachusetts General Hospital Mallinckrodt General Clinical Research Center	Massachusetts General Hospital Mallinckrodt General Clinical Research Center

Medical Practice Evaluation Center, Massachusetts General Hospital	Medical Practice Evaluation Center, Massachusetts General Hospital
Medical Practice Evaluation Center, Massachusetts General Hospital)	Medical Practice Evaluation Center, Massachusetts General Hospital)
Research Scholar Award from the Massachusetts General Hospital (MGH)	Research Scholar Award from the Massachusetts General Hospital (MGH)
Research Scholar Award from the Massachusetts General Hospital (MGH)	Research Scholar Award from the Massachusetts General Hospital (MGH)

Table A–7: Professional Association - *American Academy of Allergy, Asthma, and Immunology* - Jaccard: 80.0%

Gold Standard Cluster	**Predicted Cluster**
American Academy of Allergy, Asthma and Immunology	American Academy of Allergy, Asthma and Immunology
American Academy of Allergy, Asthma and Immunology Program	American Academy of Allergy, Asthma and Immunology Program
American Academy of Allergy, Asthma and Immunology Program Directors	American Academy of Allergy, Asthma and Immunology Program Directors
American Academy of Allergy, Asthma, and Immunology	American Academy of Allergy, Asthma, and Immunology
	Western Society of Allergy, Asthma, and Immunology

References

[1] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole Statistics-Probability Series. Taylor & Francis, 1984.

[2] Leo Breiman. Technical note: Some properties of splitting criteria. *Machine Learning*, 24(1):41–47, July 1996.

[3] Morton B. Brown. 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31(4):987–992, 1975.

[4] Peter Christen. *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Springer, 2012.

[5] Linguistic Data Consortium. Annotation guidelines for entity detection and tracking (EDT). https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-edt-v4.2.6.pdf. Version: 4.2.6 200400401.

[6] Linguistic Data Consortium. Annotation guidelines for relation detection and tracking (RDC). https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/arabic-rdc-v4.3.pdf. Version: 4.3 - 20040122.

[7] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. *Proceedings of the 23rd International Conference on Machine Learning, ACM*, 06, June 2006.

[8] Brian S. Everitt, Sabine Landau, and Morven Leese. *Cluster Analysis*. Wiley Publishing, 4th edition, 2009.

[9] Ronen Feldman and James Sanger. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, NY, USA, 2006.

[10] Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, pages 466–471. Association for Computational Linguistics, 1996.

[11] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proc. Nat. Acad. Sci.*, 46:16569, 2005.

[12] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171–1220, 06 2008.

[13] Dana J and Loewenstein G. A social science perspective on gifts to physicians from industry. *JAMA*, 290(2):252–255, 2003.

[14] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–2018. [Online].

[15] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: an introduction to cluster analysis*. Wiley, 1990.

[16] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[17] VI Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

[18] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[19] Behrang Mohit. *Named Entity Recognition*, pages 221–245. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[20] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[21] A. V. Neale, K. L. Schwartz, and M. A. Bowman. Conflict of interest: can we minimize its influence in the biomedical literature? *J Am Board Fam Pract*, 18(5):411–413, 2005.

[22] Institute of Medicine, Board on Health Sciences Policy, Education, and Practice Committee on Conflict of Interest in Medical Research. *Conflict of Interest in Medical Research, Education, and Practice.* National Academies Press, 2009.

[23] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, November 2011.

[24] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[25] Jaideep Ray, Ali Pinar, and C. Seshadhri. Are we there yet? when to stop a markov chain while generating random graphs. In *Proceedings of the 9th International Conference on Algorithms and Models for the Web Graph*, WAW'12, pages 153–164, Berlin, Heidelberg, 2012. Springer-Verlag.

[26] Peter Rousseeuw. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65, 11 1987.

[27] Dennis F. Thompson. Understanding financial conflicts of interest. *New England Journal of Medicine*, 329(8):573–576, August 1993.