Diachronic Embeddings for People in the News

Felix Hennig and **Steven R. Wilson**
University of Edinburgh
Edinburgh, UK
F.M.P.Hennig@sms.ed.ac.uk, steven.wilson@ed.ac.uk

Abstract

Previous English-language diachronic change models based on word embeddings have typically used single tokens to represent entities, including names of people. This leads to issues with both ambiguity (resulting in one embedding representing several distinct and unrelated people) and unlinked references (leading to several distinct embeddings which represent the same person). In this paper, we show that using named entity recognition and heuristic name linking steps *before* training a diachronic embedding model leads to more accurate representations of references to people, as compared to the token-only baseline. In large news corpus of articles from *The Guardian*, we provide examples of several types of analysis that can be performed using these new embeddings. Further, we show that real world events and context changes can be detected using our proposed model, with a focus on the examples of UK prime ministers and role changes in the football domain.

1 Introduction

Diachronic embeddings are an extension to traditional word embeddings that capture changes in word representations over time. These approaches have been used for several computational social science studies focused on the analysis of language and its change over time. For example, Garg et al. (2018) used embeddings to study changes in gender and racial biases over decades of literary documents, and Szymanski (2017) used diachronic embeddings to solve temporal word analogies, leading to insights about political and social changes.

In computational studies of linguistic change, news corpora have been a popular resource because of their stylistic consistency from year-to-year and the availability of large amounts of text from each individual year or even month. In the news, changes in the usage of a word often correspond to changes in the world as well, allowing for inferences about what is happening in the world, and what journalists have been focused on from changes in the embeddings over time. Previous work has also analysed the changes that representations of specific entities undergo, such as corporations like Amazon and Apple or even names of people such as *Obama* or *Trump* (Yao et al., 2018).

However, linking the surname of a person directly to a specific individual is problematic in many cases. It does not allow us to have distinct embeddings for people with the same surname, such as *Bill* and *Hillary Clinton*. A surname might also be a word, such as "may". There is *Theresa May*, the person, but also the month *May* and the verb *may*, which are often treated as the same token for the purposes of creating word embeddings. Issues like these can interfere with downstream analyses based on diachronic embeddings, leading to cases where multiple distinct embeddings exist for the same person, causing information about them to be potentially overlooked. At the same time, multiple people may be represented by the same embedding, resulting in noisy results without a clear way to determine the full set of people represented by the embedding, or a way to disentagle each person's influence.

In this work, we tackle these problems by creating explicit, diachronic embeddings for references to individual people over time, which we embed in the same diachronic space as the remaining words. We do this by finding mentions of people in texts, then linking various surface forms of the same person together and aggregating these contexts. We use named entity recognition and use full names to alleviate the problems mentioned above. All work is done on a data set consisting of more than 2 million articles from *The Guardian*, a British newspaper which has not previously been used in

Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pages 173–183

https://doi.org/10.18653/v1/P17

studies of diachronic embeddings. This provides an additional perspective to a number of studies which have exclusively focused on US-centric media outlets (Parker et al., 2011), especially the *New York Times* (Kutuzov et al., 2017; Yao et al., 2018; Szymanski, 2017). Further, we provide a set of case studies, showcasing some of the main issues with the baseline approach and the kinds of analyses that can be performed with our diachronic embeddings.[1]

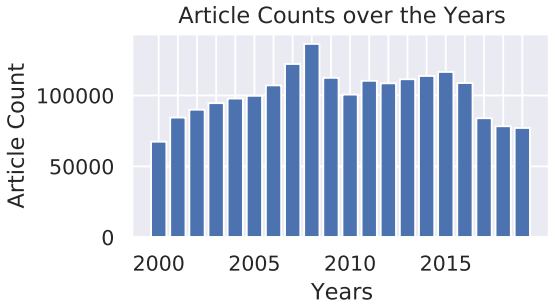

Figure 1: The article counts per year.

2 Related Work

News corpora have seen a lot of attention from researchers in diachronic embeddings to make a variety of inferences about the real world. From finding temporal analogies (Zhang et al., 2015) – such as "iPod" being the 2000s equivalent of a "Walkman" in the 1990s – observing the change of the words "amazon" and "apple" (Yao et al., 2018) or tracing armed conflicts (Kutuzov et al., 2017). Yao et al. (2018) explicitly also show examples of persons in the media changing contexts as well as the changes in the association of a role to a person (president, mayor). We use their *DynamicWord2Vec* (DW2V) model to create the diachronic embeddings. Compared to other models which are typically based on alignment of trained embeddings spaces, DW2V aligns embedding spaces during training. Kutuzov et al. (2018) provide a good overview of the field.

Kutuzov et al. (2017) do event detection instead of gradual context changes, by tracking country names in the news over time and observing state changes between war and peace. They propose the aggregation of multiple words into "concept embeddings" and manage to improve the event detection score significantly this way. This is an example of a move away from strict word embeddings towards more high-level embeddings.

Previous analysis of people in corpora as mentioned above is strictly based on the simple mapping of a token to a person, usually the persons surname is used (Szymanski, 2017; Yao et al., 2018). While this is a workable solution, we seek to show how diachronic embeddings can be improved with more explicit linking of references to people.

A similar task to linking references of people in the news for diachronic embeddings is literary character detection (Vala et al., 2015) for modeling relationships between characters over time (Chaturvedi et al., 2016). However, proposed approaches such as the Book-NLP pipeline (Bamman et al., 2014) are focused on very long texts with set of recurring characters with less ambiguous names. In news data, on the other hand, it is common for completely unrelated people to share surnames (or even full names), and the same person may be mentioned in a large number of separate documents.

3 Data

We use a data set sourced from the British newspaper *The Guardian*. *The Guardian* provides all their content via an API called OpenPlatform[2], launched in 2009 (Anderson, 2009). This data source has seen only tangential use in the scientific community (Li et al., 2016; Guimarães and Figueira, 2017; Murukannaiah et al., 2017) and has not been used for diachronic models before.

The data set contains a total of $2,021,947$ articles spread over the years 2000 to 2019, containing 1.65 billion tokens. The documents[3] were retrieved from the API in March 2020. Each document consists of the article body and additional meta-information such as the section in which the article was published. We used spaCy[4] to tokenize the text, and the resulting data was a collection of token sequences, divided into yearly chunks.

Figure 1 shows the distribution of articles over the years, each year contains about $100,000$ articles, which is about 270 articles every day. The distribution of articles over sections is shown in Figure 2. Articles are distributed very broadly over topics ranging from media over sports to politics.

[1] Our pre-trained diachronic embeddings as well as the code to generate them are published at github.com/fhennig/DiachronicPeopleEmbeddings

[2] open-platform.theguardian.com

[3] Only documents with the types `article` and `liveblog` were included; this excludes other types of content like crossword puzzles.

[4] spacy.io

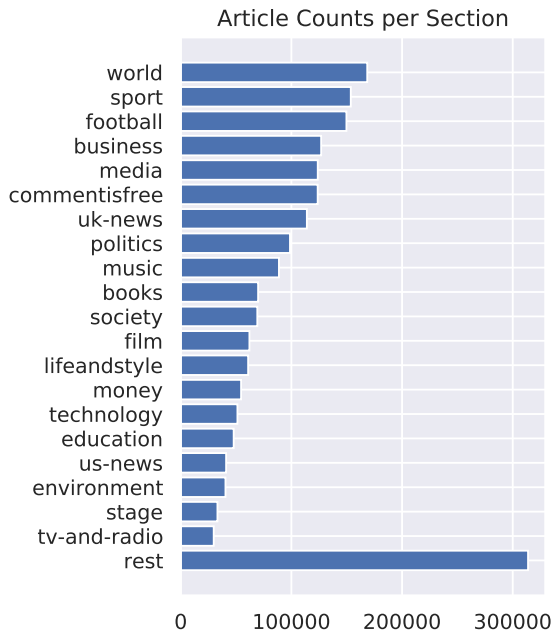

Figure 2: The article counts per section.

Notably, football is a distinct section from sports and contains about the same amount of articles.

4 Modelling People in the Text

While it is common for a person to be only referred to by their surname, a reference to a person in a text can take different forms, which should be linked to a single entity to allow the creation of consistent embeddings for people's names. Once a set of mentions have been linked to a person, a new text corpus is built where any span of tokens that is part of a mention is replaced by a pseudo-token representing the person.

4.1 Identifying Mentions

The baseline method for identifying the mention of a person in a text is to look for their surname, a single token. However, a person can also be referred to by their name, role or with co-references. We focus only on mentions by name, but go beyond the single-token method. We identify all occurrences of full or partial names and link them together based on context.

We identify names in the text using named entity recognition (NER) implemented in spaCy. It provides a neural NER model with an architecture based on Strubell et al. (2017), we used the default pretrained model for English: `en_core_web_sm`. The library creators report an accuracy of around 85% [5] for all entities, we only use the detection of persons for which no individual score is given.

The model detects full names (*Tony Blair*), partial names (*Blair*) mentions with title (*Mr. Blair*) and possessive mentions (*Blair's*). For subsequent linking of mentions, the possessive "'s" is ignored.

4.2 Merging Mentions

The subsequent linking of mentions relies on the structuring of the corpus into distinct documents, as well as the temporal structure to link local mentions more leniently – within an article or within a fixed time span. In writing, using references and shortening of names relies on saliency of the name to the reader, this is partially mimicked by the steps taken below.

Within an article, a person is often introduced by their full name, and subsequently only a part of their name – often the surname – is used to refer to them. In that case, the surname is non-ambiguously referring to the person identified by the full name in that same article. Therefore, within an article, we link any detected mention whose tokens are a subset of a previously seen mention in that article to that previous mention.

Across articles, mentions are then linked together based on exact match of their name, which typically means first and last name.

We observed that for very well-known people, there were quite a few articles which referred to them only by their surname, never mentioning their full name in the article. This is typically done for presidents and prime ministers, which are famous enough to be associated with their surname only. To merge these occurrences, for every month, every name that consists only of a single token is linked to the person whose name contains that token and has the most mentions in that month and the previous month. The same is done for names that start with *Mr* or *Mrs*. This helps mapping *Mr Johnson* and *Johnson* to *Boris Johnson*. This also produces a few false positives, such as *Mrs Blair* getting mapped to *Tony Blair*. Doing this matching at the month-level allows the ambiguous reference *Clinton* to be mapped to *Bill Clinton* in one time period and *Hillary Clinton* in another.

Once linking is complete, for every mention the span of tokens forming the mentions is replaced with a single token: the full unique name of the person that is referenced.

[5] spacy.io/usage/facts-figures#spacy-models

Mentions	People
> 5	458,158
> 50	61,828
> 500	7,559
$> 5{,}000$	417
$> 50{,}000$	10

(a) Number of people that have more than a specific number of mentions, illustrating a logarithmic relationship.

Person	Count
Tony Blair	144,061
Donald Trump	143,149
David Cameron	119,176
Gordon Brown	116,895
Boris Johnson	84,744
Hillary Clinton	67,700
Chelsea	62,336
Commons	58,079
George Bush	58,018
George Osborne	56,646

(b) The 10 people with more than 50,000 mentions.

Table 1: The distribution of people and mentions, as well as the top 10 people by mention count.

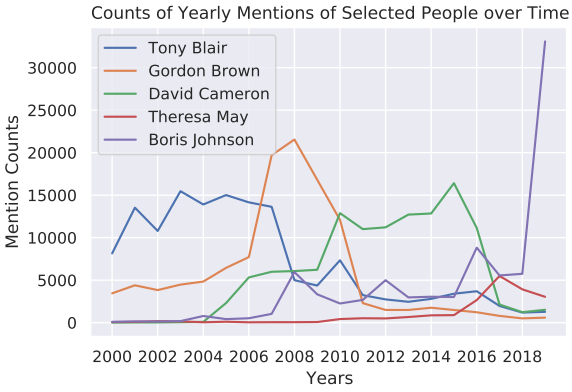

Figure 3: Mentions of the 5 prime ministers of the UK from 2000 to 2019.

4.3 Data Analysis

The identifying and linking of occurrences of people in the corpus allows for an initial analysis and provided useful insights for subsequent experiments with embedding models.

Persons in the Data Set Overall, there are 2,725,110 persons with 29,943,111 total mentions in our dataset.[6] Table 1a shows how many people have how many mentions. The distribution is logarithmic; like vocabularies of corpora, the distribution of people in the corpus follows Zipf's Law. A few people are mentioned very frequently, while most people are only referred to a few times.

Table 1b shows the 10 people that have at least 50,000 mentions. *Chelsea* and *Commons* are notable false positives; the quality of the name detection is discussed below. The other 8 people are all politicians: four from the UK and four from the US. In fact, the top 100 mentions are dominated by politicians, but athletes, in particular football players, make up a large portion, too.

The raw frequencies of mentions can be analysed to make approximate inferences about the world at a given point in time. Figure 3 shows the mentions of the prime ministers of the UK for the past two decades. The mentions show popularity trends corresponding to their terms as prime ministers. For *Tony Blair*, *Gordon Brown* and *David Cameron* these curves give a good overview of when they were in office, but after 2016 it is not entirely clear who would be prime minister; from the frequencies it does not look as if Theresa May would ever be prime minister. We further analyse the prime ministers using our diachronic models in Section 5.2.2.

Quality of NER As mentioned above, the NER used has a reported accuracy of around 85%. "Facebook", "Twitter" and "Brexit" were falsely identified as persons, just like the previously mentioned "Commons" – likely detected as a name in the phrase "House of Commons", a British political institution. Besides the already reported "Chelsea", there are also "Tottenham", "Manchester United", "Fulham" and many other sports teams in the false positives. For the creation of the embeddings, these false positives are not an issue, because they are detected consistently.

There were also many instances of false positive detected names of the form "Manchester 1 - 1 Norwich", a match result. In this case, a detected mention like this cannot be merged and "ties up" the words "Manchester" and "Norwich", preventing them from contributing to any embeddings. In our investigation of a sample of these embeddings, however, we observed that this effect was not significant. Tokens such as "Manchester" are typically very common throughout the corpus and so we can still build reliable embeddings for them, even if some of their occurrences are missed due to being treated as names.

Duplicate and Ambiguous Names Our modelling of mentions was motivated by problems with the token based surname to person mapping, namely the prevalence of duplicate and ambiguous surnames. Using the detected and linked mentions as described above, we can assess the prevalence of both phenomena.

We evaluate duplicates year by year. Each year

[6] In order to avoid the inflation of counts of a person's name due to false positives in the linking process, only exact matching was used to link mentions *across articles* for the results presented in this section only (as opposed to heuristically linking surname-only mentions *across articles* as well). *Within-article* linking was still performed as usual.

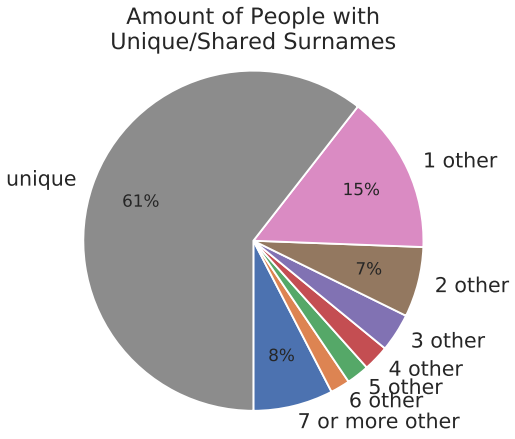

Figure 4: Percentages of people with shared/unique surnames. For each year, every person with at least 50 mentions was considered. Only names within the same year were used to identify duplicates.

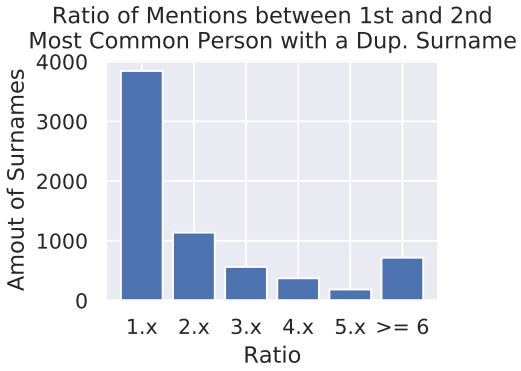

Figure 5: The plot shows the relative frequency of people that share a surname. I.e. if two people share the name *Johnson*, how many times more is the more common *Johnson* mentioned compared to the other?

we retrieve any person with at least 50 mentions and group them by surname. For every surname we then get a count of how many people share this surname. Figure 4 shows the aggregated distribution of these counts. More than a third of people do not have a unique surname.

For people with the same surname, we were also interested in whether they would all be mentioned approximately the same amount or if it is common for one person to dominate the total mentions of the surname.

Figure 5 shows the ratio of mentions between the most mentioned person with a surname and the second most. We observe that it is not uncommon for a single person to "dominate" a surname, dwarfing any mentions of other people with the same surname. Although in most cases, the first and second most commonly mentioned person with a shared surname do not have a big difference in counts of mentions.

Some surnames are also proper nouns or adjectives, we call these *ambiguous* names. The prevalence of these names is difficult to quantify due to false positives in the NER, but in 2019, from all people with at least 200 mentions, *Philip Green*, *Arron Banks* and *Fiona Hill* have the surnames that are most likely to also be used as a regular word. In all three cases, their surnames is 50 times more likely to be used as a regular word than as their surname.

In token based models, both phenomena – duplicate and ambiguous surnames – can lead to a person being "invisible", because any individual part of their name is too common on its own in another context. In Section 5.2.1 we present an embedding example for each phenomenon.

5 Diachronic Model

Once persons have been identified and linked, we trained diachronic models with the *Dynamic-Word2Vec* (DW2V) model (Yao et al., 2018). The model learns embeddings for all time slices concurrently, encoding within-time-slice similarities of tokens as well as inter-time-slice similarities of tokens to themselves at the same time. This eliminates a subsequent alignment step and also makes embeddings more stable.

The experiments and analysis of the models investigate the following questions both qualitatively and quantitatively: (1) How well can the explicit person model generate traces for people, compared to a token based baseline? (2) How well do the traces model context changes in the real world?

5.1 Experimental Setup

The baseline model uses no person detection and is purely token based; subsequently called the `token` model. Our model – subsequently called the `person` model – embeds detected persons using within-article and cross-article merging as described in Section 4.

Both models use yearly slices, covering 20 years from 2000 to 2019 (inclusive). To define the vocabulary that is used, only words or names with more than 500 occurrences across the whole time span were considered. The vocabularies of both models differ, but both contain around 55,000 types. Many

Model: Year	person *Taylor Swift*	token taylor	token swift	token johnson	person *Boris Johnson*	person *Dustin Johnson*
2010	Iggy Pop	adam	immediate	davies	mayor	Francesco Molinari
2011	Selena Gomez	davies	swiftly	alex	Bravo Boris	mcilroy
2012	Patti Smith	adam	speedy	ryan	Ed Miliband	Charl Schwartzel
2013	Beyoncé	craig	timing	nick	Iain Duncan Smith	Zach Johnson
2014	Lily Allen	jones	kim	joe	Iain Duncan Smith	Jim Furyk
2015	Madonna	smith	adele	adam	vince	Jason Day
2016	Justin Bieber	adam	beyoncé	tony	David Cameron	Jason Day
2017	Beyoncé	smith	recall	cameron	Theresa May	Jordan Spieth
2018	Beyoncé	mitchell	swiftly	jeremy	Jeremy Corbyn	Brooks Koepka
2019	Madonna	ross	drake	boris	Jeremy Corbyn	Phil Mickelson

Table 2: The table shows the closest words for different people/tokens in the two models, over the last 10 years. *Taylor Swift* is shown in the person model, and her first and last name as tokens in the baseline model, showing the ambiguity of the names by themselves. Similarly, the name *Johnson* is shown in the baseline model, and two specific Johnsons are shown in our person model.

common words are in both vocabularies, most differences are in the names, which appear as individual tokens in the token model and as compound names in the person model. Further details about the model training can be found in Appendix A.

5.2 Qualitative Analysis

Here, we first exemplify the problem of ambiguous and duplicate names, then show the modelling of the role of UK prime minister as an example of real world change represented in the model.

5.2.1 Taylor Swift, Boris and Dustin Johnson

Table 2 shows the closest word every year for the last 10 years, for the people *Taylor Swift*, *Boris Johnson* and *Dustin Johnson* in our person model, as well as the names taylor, swift and johnson as tokens in the baseline model.

Taylor Swift is an American pop musician, and provides an example of an ambiguous name that is difficult to find in the baseline model. In most years from 2010-2019, neither her first nor last name is associated with meaningful tokens in the baseline model; taylor is in the vicinity of various common names, and swift is mostly associated with other words related to the meaning "happening quickly or without delay". Our model places *Taylor Swift* into the neighborhood of other pop musicians throughout the years, such as Beyoncé and Madonna.

The name *Johnson* is provided as an example of a duplicate surname. In the baseline model, the surname is associated with various first and last names throughout the first 7 years of the shown time frame. In the last 3 years, associated tokens indicate *David Cameron* and *Jeremy Corbyn* – two British politicians, as well as the first name of *Boris Johnson*. In these years, *Boris Johnson* was by far the most mentioned person with his surname. Looking at his neighboring words and people in our model, we see him associated with other politicians throughout the years or with his role ("mayor"). *Dustin Johnson* also shares the surname, he is an American golfer with a relatively small amount of mentions throughout the years. The baseline model does not show him at all, while our model associates him with other golfers, even people he played with (i.e. *Jordan Spieth* in 2017).

5.2.2 Prime Ministers of the UK

We present the change of the UK prime minister from 2000-2019 as an example of the representation of real world change in the model. We use the the vector of the incumbent prime minister in 2010 – the middle of the time range – as the vector for the role "prime minister". For all other years we retrieve the closest person or token to this vector. Table 3 shows the closest person/token in our new model and the baseline. Our model is mistaken only twice, while the baseline is wrong 8 times. Furthermore, the token model does not allow searching only for people and it is difficult to know how to associate tokens with names.

The biggest differences are seen for *Gordon Brown*, *Theresa May* and *Boris Johnson*. The token model does not retrieve *May* or *Johnson* at all, and retrieves *Gordon Brown* only one out of 3 years, and with his first instead of last name. From the neighborhood of the tokens in the embedding space we found that brown is predominantly associated with the color and may with its usage as a modal verb. Johnson is not ambiguous, but a very common surname.

Year	Our Model	baseline
2000	Tony Blair	hague (blair)
2001	Tony Blair	chancellor (blair)
2002	Tony Blair	blair
2003	Tony Blair	blair
2004	Michael Howard*	chancellor (blair)
2005	Tony Blair	chancellor (gordon)*
2006	Tony Blair	chancellor (gordon)*
2007	Gordon Brown	blair*
2008	Gordon Brown	gordon
2009	Gordon Brown	cameron*
2010	David Cameron	cameron
2011	David Cameron	cameron
2012	David Cameron	cameron
2013	David Cameron	cameron
2014	David Cameron	cameron
2015	Jeremy Corbyn*	cameron
2016	Theresa May	cameron*
2017	Theresa May	jeremy*
2018	Theresa May	jeremy*
2019	Boris Johnson	jeremy*

Table 3: The table shows the names/tokens closest to the vector for *David Cameron* and `cameron` respectively, in 2010. For the token based model, the token in parenthesis is the closest name, shown for a fairer comparison. The asterisk indicates incorrect associations.

5.2.3 Largest Change Spikes

Simple factual information such as a person's role can be found with the model, while spikes in the model can help us to detect real world *events*. The largest context changes in the model were identified by calculating cosine distance scores for every word to itself throughout the years. Inspection of samples from the largest spikes showed that these detected spikes corresponded to long term change events as well as one-off events. Examples for long term change events that we observed are the election of *Imran Khan* as prime minister of Pakistan in 2018 or the Scottish football player *David Marshall* joining the national team in 2004. *Imran Khan* also attended a series of sports events in 2015, which also showed as a significant context change in the model. However in this case it was only a one-off event. In general, career changes, promotions and team changes for athletes (long term changes) as well as accidents, wins of competitions or legal disputes (one-off events) were frequently observed causes of context change.

We also noticed that despite using the full name of a person, there are still duplicates that are falsely linked together. This happens when a person has a common first and last name, such as *David Marshall* or *Scott Walker*. Sporadic reporting about two different people in different domains then appears as a context change of a single person.

Overall, spikes are most pronounced in medium to low data settings, where just a few articles can already have a big influence on the context of a person. With a lot of articles, a person's context fluctuates less, but is more precise.

Examining surnames in the baseline model showed that uncommon first or last names still showed similar spikes as the `person` model, but many people could not be found at all in the baseline model due to their names being too common.

5.3 Quantitative Analysis

Quantitative analysis is a difficult task with diachronic models because there are no accepted gold standard embedding spaces to compare against. Yao et al. (2018) use the sections associated with the articles to group words together. They then identify clusters in the embedding space and compare them with the groupings created by the sections. Kutuzov et al. (2017) track countries in the news and use a manually created database of armed conflicts as a target to compare their model against.

We present two experimental setups, one following Yao et al. (2018) based on the sections of articles and one inspired by the change event detection by Kutuzov et al. (2017), comparing our model against gold standard change events in the football domain.

5.3.1 Section Analysis for Ambiguous Names

One way to measure the quality of embeddings in bulk is to assess the overlap of clusters in embedding space with clusters given by the sections assigned to each article in the corpus (Yao et al., 2018). Section labels were not used in the training process, and are semantically cohesive, so we expect that people from the same section should cluster together in embedding space, too.

We derive the dominant section of people per year by finding the section with the highest count of mentions for every person for that year. Persons with a dominant section that has less than 35% of their total mentions were not included due to the section association being too unclear (following Yao et al. (2018)). Every yearly person vector was also only included if the person had at least 100 mentions that year, to ensure some degree of stability in the embeddings (Burdick et al., 2018). Note that the sizes of sections by article count are heavily imbalanced. This translates also to the number of people in the sections: the number of people in a section can differ by two orders of magnitude.

Cl.	yearly		total	
	t. model	p. model	t. model	p. model
non-ambiguous names only				
10	0.6425	**0.6484**	0.6575	**0.6780**
15	0.6565	**0.6731**	0.6602	**0.6671**
20	0.6506	**0.6665**	0.6367	**0.6632**
25	0.6374	**0.6575**	0.6544	**0.6556**
ambiguous names included				
10	0.6472	**0.6478**	**0.6641**	0.6261
15	0.6559	**0.6698**	0.6588	**0.6748**
20	0.6501	**0.6645**	0.6537	**0.6785**
25	0.6400	**0.6586**	0.6523	**0.6630**

Table 4: Normalised mutual information scores for various experimental settings. The Cl. column represents the number of clusters. The "yearly" columns contains averaged clustering results by year. The "total" column contains clustering results of all vectors across all years in one space.

To make our model and the baseline directly comparable, full names had to be mapped to tokens in the baseline model; we used the surname of each person as their token. For a fair comparison, the mapping had to be injective, so only people with a unique[7] surname were included.

The vectors from all years can be clustered in a single vector space, or clustering can be done for each year and the results averaged over years. The full space emphasises cross-year alignment, whereas the yearly clustering emphasises local separation. We report results for both as `total` and `yearly`. For both settings, we always create three clusterings and average the results to smooth out effects of random initialisation. We used spherical k-means[8] to create the clusters. Two sets were evaluated:

(1) `non-ambiguous only`: For every mapped token, we ensured that the mapped token *also* appeared at least 35% of the time in the same section as the full name (15,622 vectors).

(2) `ambiguous names included`: All the names are included, regardless of the distribution of the associated token (20,084 vectors).

As an example, in 2019 the dominating section for *Theresa May* was `politics` with 63% prevalence. The word `may` appeared mostly in `politics` too, but only with 8% prevalence. Therefore, *Theresa May* is included in the second set, but not in the first.

Table 4 shows the results of the experiments. Across both data sets and both experimental setups, for various cluster sizes, the `person` model outperforms the `token` model except in the 10 cluster setup with the ambiguous names included.

It is interesting to see that the gains in the set containing the ambiguous names are not larger, as it would be expected. A potential reason could be that while, for example, `may` appears frequently in many sections, the generic contexts "cancel out", leaving the political context given by *Theresa May* as the dominant context for clustering, even though the immediate neighbourhood of the word does not contain words indicating that.

5.3.2 Football

As described above, context changes of a person show up as vector space movement in the model. We quantify this using football players that change their role from player to coach. This role change is an important event and there is a lot of reporting on the football domain, with detailed instead of broad reporting.

We use Wikidata (Vrandečić and Krötzsch, 2014) as the source for career change information. Wikidata is an open access, community maintained knowledge graph containing over 80 million nodes. To retrieve the relevant persons, we first selected people by name and then filtered the list based on specific properties to eliminate duplicates. Out of a list of names appearing in the football section of the corpus, 39 players were retrieved that became coaches in the years between 2001 and 2019. The full list can be found in appendix B.

Based on the assumption that their change from being a player to becoming a coach was their biggest change in their context, we looked for the biggest change in the embedding space throughout all years, in the `person` model. For 8 out of 39 people, their biggest change spike coincided with the year in which they were first a trainer. This gives a 21% accuracy, 4 times better than random guessing. For 5 additional people, during the year of their career change, a spike occurred that was larger than 1 standard deviation from the mean change.

6 Discussion of Results

In the analysis of the persons that appear in the corpus we showed the extend to which duplicate names are prevalent, showing that more than a third of people do not have a unique surname. In the examples from the embeddings we showed how these

[7] As we only included people with over 100 mentions, some non-unique surnames may be considered unique if only one person with the surname has more than 100 mentions.

[8] We use the implementation provided by the spherecluster Python package.

ambiguous and duplicated surnames prevent meaningful embeddings for these names in the baseline model. With the example of the UK prime minister we showed that our model identified the prime minister each year with 90% accuracy compared to 65% for the baseline model.

For people with unique but potentially ambiguous surnames, our model improved embeddings as well. Improvements were minor, but consistent across 15 out of 16 different experiments.

For the analysis of the change tracking in the model, the case of the prime minister of the UK showed that the explicit person embeddings improved the association markedly. A systematic look at change detection in role changes for football players showed performance four times better than the random baseline, but there is still room for improvement.

Context changes in the model can be linked to real world events, but not all real world changes appear in the model. This may be explained by selective reporting in the source corpus, as not all events that actually happen are reported on. Reporting is also selective in other ways: for a role change like an election and a new prime minister, the reporting focuses more on the new person overtaking the role than on the person being replaced. When a prime minister changes, there is a much more noticeable context change for the person getting into office than the person getting out of office. For athletes, team changes or role changes are also less easy to detect than complete out of context reporting, about for example sexual misconduct or drunk driving.

Overall, the detection and linking of names worked well and was a large improvement over using surnames only. It enabled us to identify certain people in the first place and disambiguate people with common surnames. The NER has a few false positives, but false positives are consistent (i.e. Twitter is misclassified as a person, but in every occurrence).

7 Conclusion

We have shown that full name identification of individuals in the news based on NER and heuristic linking is doable and provides meaningful insights linkable to real world developments. This approach is a simple yet effective improvement over a token based baseline. We quantified some common problems with the baseline approach, such as duplicate and ambiguous surnames. We provided some initial analysis of event detection both based on samples and a small quantitative experiment, showing promising results, and we provided an analysis of a previously understudied, British news corpus.

Future work on the model can incorporate improvements to NER models as well as expanding the detection of references to non-name references such as role references, co-references. Linking can be improved with a more sophisticated treatment of name variants and titles. These steps have been shown to improve character detection in literary texts (Vala et al., 2015). Lastly, gaps in reporting could be filled by relying on more than a single newspaper as a source of text.

References

Kevin Anderson. 2009. Guardian launches Open Platform tool to make online content available free. *The Guardian*.

David Bamman, Ted Underwood, and Noah A Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379.

Laura Burdick, Jonathan K Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Nuno Ricardo Pinheiro da Silva Guimarães and Álvaro Pedro de Barros Borges Reis Figueira. 2017. Building a Semi-Supervised Dataset to Train Journalistic Relevance Detection Models. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pages 1271–1277.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2017. Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the Events and Stories in the News Workshop*, pages 31–36, Vancouver, Canada. Association for Computational Linguistics.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The Role of Discourse Units in Near-Extractive Summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.

Pradeep K. Murukannaiah, Chinmaya Dabral, Karthik Sheshadri, Esha Sharma, and Jessica Staddon. 2017. Learning a Privacy Incidents Database. In *Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp*, HoTSoS, pages 35–44, Hanover, MD, USA. Association for Computing Machinery.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition LDC2011T07. Technical report, Linguistic Data Consortium.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. *arXiv:1702.02098 [cs]*.

Terrence Szymanski. 2017. Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453, Vancouver, Canada. Association for Computational Linguistics.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic Word Embeddings for Evolving Semantic Discovery. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*, pages 673–681.

Yating Zhang, Adam Jatowt, Sourav Bhowmick, and Katsumi Tanaka. 2015. Omnia Mutantur, Nihil Interit: Connecting Past with Present by Finding Corresponding Terms across Time. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 645–655, Beijing, China. Association for Computational Linguistics.

Appendix

A Diachronic Model Implementation Details

For the DW2V model Yao et al. (2018) provide their hyper-parameter settings as a starting point: $\lambda = 10, \tau = \gamma = 50$ and window size 5 for the PPMI matrices, 5 epochs of training. The absence of gold standard embeddings or other high-quality target data hinders a systematic hyper-parameter search. We briefly trained some models using slightly varied parameters to adjust to the different data size, but no large improvements could be found. The model creates two embeddings for each word, which we concatenate to form the final word embeddings. In comparison to the reference implementation, both the batches as well as the order in which the time slices are updated are randomised, to prevent any skewed embeddings that would be created by a fixed training order.

B Football Players and Coaches

2001	2014
Walter Mazzarri	Paul Scholes
Didier Deschamps	2015
Roberto Mancini	Jürgen Klopp
2003	David Wagner
Massimiliano Allegri	2016
2004	Zinedine Zidane
Ian Rush	Patrick Vieira
2005	Unai Emery
Henning Berg	Olof Mellberg
Paul Gascoigne	2017
2006	Harry Kewell
Gareth Southgate	2018
Antonio Conte	Thierry Henry
Diego Simeone	Marco Silva
2007	Joey Barton
Thomas Tuchel	Sol Campbell
Pep Guardiola	Garry Monk
Paul Le Guen	2019
2008	Jonathan Woodgate
Luis Enrique	Jürgen Klinsmann
2009	Scott Parker
Jaap Stam	Duncan Ferguson
Vincenzo Montella	Dick Advocaat
Mauricio Pochettino	Mikel Arteta
2011	Pepe Mel
Dietmar Hamann	
2013	
Laurent Blanc	

Table 5: Players that became coaches between 2001 and 2019

Table 5 shows the full list of players who later became coaches between 2001 and 2019, as retrieved from Wikidata. The analysis of these people was presented in section 5.3.2.