Extracting Networks of People and Places from Literary Texts

John Lee and Chak Yan Yeung
Halliday Centre for Intelligent Applications of Language Studies
Department of Chinese, Translation and Linguistics
City University of Hong Kong
{jsylee,chayeung}@cityu.edu.hk

Abstract

We describe a method to automatically extract social networks from literary texts. Similar to those in prior research, nodes represent characters found in the texts; edges connect them to other characters with whom they interact, and also display sentences describing their interactions. Furthermore, other nodes encode places and are connected to characters who were active there. Thus, these networks present an overview of the "who", "what", and "where" in large text corpora, visualizing associations between people and places.

1 Introduction

To fully understand a matter, one must be able to answer, as it were, the "Five W" questions: *who*, *what*, *where*, *when*, and *why*. In Humanities research, scholars comb texts to answer similar questions --- who the principal figures were, with whom they interacted, what they did, where and when they lived, and why they made an impact. The vast amount of texts available in digital libraries has, on the one hand, enlarged the breadth on which scholars can perform textual research (Crane, 2006); on the other hand, the sheer volume overwhelms an individual's ability to read the texts in depth to answer these questions.

Overviews — information abstracted from a collection of texts — can help a reader rapidly grasp the scope and nature of the collection in question (Greene et al., 2000), thereby supporting "distant reading" of large text corpora (Moretti, 1999). Ideally, they should also serve as gateways to the primary source by helping the reader locate points of interest for closer reading.

Manually written overviews tend to be centered on one of the W's. For example, biographies summarize the "who" in a text; a plot précis explains the "what" of a novel; and a gazetteer gives a list of locations. Most approaches in computational linguistics also focused on each of the W's in isolation. Named entity recognition systems retrieve lists of personal entities, organizations, geographical names, and the like (Chinchor et al., 1999); temporal resolution systems detect temporal expressions (Mani and Wilson, 2000); discourse parsers can help answer *why* questions (Marcu, 1998).

In more recent work, there has been much effort to synthesize two or more of the W's, for example, detecting co-occurrences of dates and place names (Smith, 2002); linking time to events (Pustejovsky et al., 2005); connecting people to the events in which they interact with others (Doddington et al., 2004; Agarwal et al., 2010); as well as "nexus points" of groups of people at particular locations (Bingenheimer et al., 2009). This paper contributes another step in this direction, reporting the first attempt to automatically construct social networks from literary texts integrating *who*, *what*, and *where*.

The rest of the paper is organized as follows. The next section reviews previous work in the automatic generation of social networks. Section 3 defines the research question. Section 4 describes the baseline and our generation algorithm. Sections 5 and 6 outline our data and evaluation results. The paper concludes with future work in the last section.

2 Previous Work

2.1 Conversational networks

Most research in automatic generation of social networks has concentrated on extracting the "who" and the "what" from a corpus. More precisely speaking, these networks should be termed "*conversational networks*." Typically, they consist of nodes representing people, and directed edges encoding the nature of their communication. The earliest attempts are concerned with structured corpora, where the senders and receivers of such communications are clearly defined, such as in internet relay chat (Mutton, 2004) and e-mail messages (Diesner et al., 2005). The edges contain analyses of the content of the messages, such as the topics and the words used.

Likewise, when applied on literary texts, automatic generation of social networks has also focused on dialogues between characters. For example, in networks constructed from Shakespearean plays, two characters are considered connected if one is speaking and the other is also on stage (Stiller et al., 2003). The edge can also characterize the speech, for example the distribution of verb tense and person in networks of Classical Greek tragedies (Rydberg-Cox, 2011). For novels, dialogues between characters are not explicitly stated, and must be identified using techniques in quoted speech attribution. A conversational network can then be similarly built; the edges can characterize, for example, the length of dialogues between the two characters (Elson et al., 2010).

2.2 Social Networks

Relations between people, however, are not described only, or even primarily, by conversations, in most other genres. The Automated Content Extraction (ACE) task, which focuses on newswire text, aims to infer all entities mentioned in a text, the relations among them, and the events in which they participate (Doddington et al., 2004). Also using newswire corpora, Agarwal and Rambow (2010) extract social events using features from syntactic parse trees. Emphasizing the cognitive states of the participants, they classify the events into "interactions" or "observations". In the extraction of social networks from biographies, personal relationships are classified as "positive" or "negative" (van de Camp and van den Bosch, 2011).

Our goal is to produce overviews of large corpora of literary texts, and is thus most similar to that of (Elson et al., 2010). Our networks are not, however, limited to conversations, so that quoted speech needs not be assumed to be the main vehicle of encoding interpersonal relations; in this sense, our scope is closer to (Agarwal and Rambow, 2010). Besides people and their associated events, our networks also integrate locations. Whereas past research have focused on toponym resolution, i.e. linking place names to geographical coordinates (Smith and Crane, 2001; Speriosu et al., 2010), we attempt to link them to events in the text. In summary, this paper is the first attempt to extract beyond conversational networks from literary texts, and encompass not only *who*, but also *what* and *where*.

3 Research Question

For texts that are rich in dialogue interactions, such as novels and serials, social interactions can be well represented by conversational networks (Elson et al., 2010). Such networks are less suitable for texts in most other genres, where evidence concerning the characters' social relationships is found largely outside of dialogue interactions. For example, in the book of *Genesis*, the tense relationship between Sarai, Abram's wife, and Hagar, Sarai's servant, is mentioned frequently, but the two of them are never involved in any dialogue interactions in the book. In fact, there are 330 distinct personal names in *Genesis*, but only 53 are involved in any dialogue interactions, so the above method would only be able to capture the social relationships of one-sixth of the total characters.

An alternative method, therefore, is needed to extract social networks from texts that lack dialogue interactions. We now define the structure (Section 3.1) and meaning (Section 3.2) of the networks to be generated, then describe our proposed method (Section 4).

3.1 Network definition

Our network graphs contain two types of nodes, one encoding people ("who"), and the other encoding locations ("where"). Each personal name is presented as a node (a 'person-node'). Two person-nodes are connected by an edge (a 'person-person edge') if there is textual evidence, i.e. a set of sentences in the corpus attesting that the two people are kin or have at least one instance of social interaction, as defined in Section 3.2.

Since some social relationships do not occur in any geographical context, and some span over multiple locations, we decided to treat the geographical names as another type of nodes ('location-nodes'), rather than attaching them to the person-person edges. A person-node and a location-node are connected by an edge (a 'person-location edge') if the person has been to that location physically.

In both person-person and person-location edges, we encode the source text that supports the claim ("what"). This design allows the readers to see the relationships of each person and the activities in each location easily. Figure 1 shows an example social network graph.

3.2 Network Construction

Person-Person Edges: As pointed out by Agarwal and Rambow (2010), a text may describe social relations between two people *explicitly* or *implicitly*.

Explicit descriptions typically state the relationship, e.g., kinship, between two people. Consider the sentence "[Noah] had three sons: [Shem], [Ham], and [Japheth]." The father and son relationships (Noah - Shem, Noah - Ham and Noah - Japheth) are explicitly mentioned, but the sibling relationships (Shem - Ham, Shem - Japheth and Ham - Japheth) can also be inferred. Our practice is to annotate the former, but not the latter type of relationships.

Implicit descriptions, in contrast, "create or perpetuate a social relationship" between two people through an event. In our annotations, events can be verbal or non-verbal interactions.

Verbal interactions. Two people are said to have a verbal interaction when one or both of them speaks, and both are aware of the communication. This type of interaction may be either quoted speech[1], or communications that are implied but not presented in the text as actual dialogues[2].

Non-verbal interactions. Two people are said to have a non-verbal interaction when they interact non-verbally and are mutually aware of the interaction. This type of interaction may involve direct physical contact between the people[3], non-physical contact[4], and others which are ambiguous due to lack of detail[5].

For each relation, the words in the sentence that indicate that relation are also annotated. For implicit descriptions, the majority of these are verbs. For example, the word 'treated' in the sentence 'Sarai treated Hagar harshly' was extracted. For explicit descriptions, these are mostly nouns, e.g., 'son'.

Person-Location Edges: An edge is placed between a person-node and a location-node if it can be inferred from the text that the person has physically been to that location. For example, based on the sentence "[Esau] went to [Ishmael] and married [Mahalath]", both Esau and Mahalath are connected to the place Ishmael.

4 Proposed Approach

We first describe our baseline (Section 4.1); then our proposed algorithm, incorporating coreference (Section 4.2), syntactic and semantic information (Section 4.3); and finally a second baseline using a machine learning approach (Section 4.4).

4.1 Baseline

It is assumed that the input text already has its personal and geographical names marked up, either manually or with a named entity recognizer. For social relationships stated outside of dialogue interactions, the named entities may be expected to be in relatively close proximity to each other. Our baseline is therefore co-occurrence: any two personal names that co-occur in a sentence are connected in the graph. Likewise, any personal name and geographical name that co-occurred were also connected.

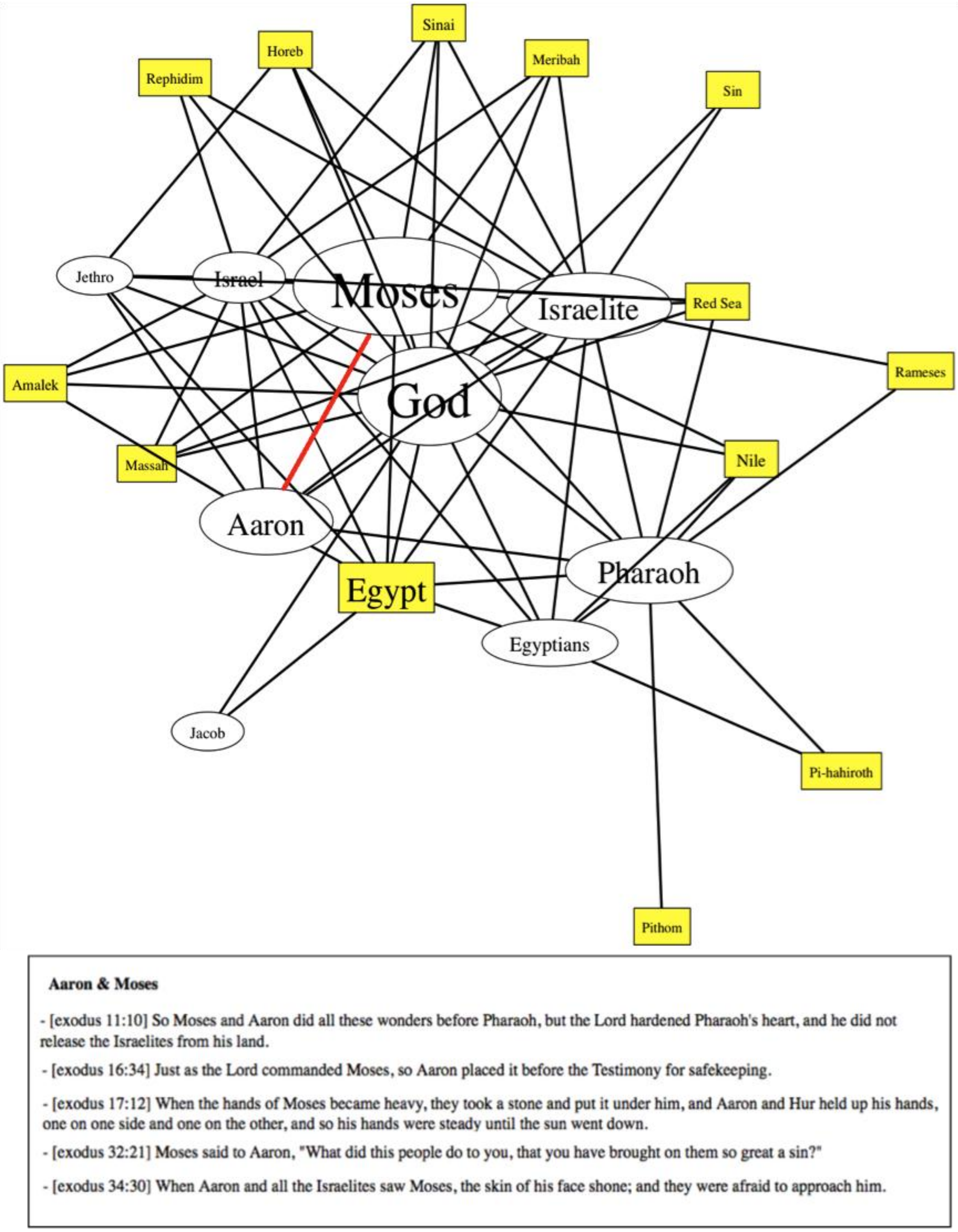
[1] E.g., "[Esau] said, "I have plenty, my brother. Keep what belongs to you." "No, please take them," [Jacob] said."

[2] E.g., "[Isaac] spoke to his son [Esau]."

[3] E.g., "While they were in the field, [Cain] attacked his brother [Abel]"

[4] E.g., "[Enoch] walked with [God] for 300 years."

[5] E.g., "[Sarai] treated [Hagar] harshly."



Aaron & Moses

- [exodus 11:10] So Moses and Aaron did all these wonders before Pharaoh, but the Lord hardened Pharaoh's heart, and he did not release the Israelites from his land.

- [exodus 16:34] Just as the Lord commanded Moses, so Aaron placed it before the Testimony for safekeeping.

- [exodus 17:12] When the hands of Moses became heavy, they took a stone and put it under him, and Aaron and Hur held up his hands, one on one side and one on the other, and so his hands were steady until the sun went down.

- [exodus 32:21] Moses said to Aaron, "What did this people do to you, that you have brought on them so great a sin?"

- [exodus 34:30] When Aaron and all the Israelites saw Moses, the skin of his face shone; and they were afraid to approach him.

Figure 1: A portion of social network drawn automatically from *Exodus*, the second book in our test set. Person-nodes are circular in shape and location-nodes are rectangular in shape. Some of the sentences associated with the selected edge are displayed at the bottom. The more frequently a name is mentioned in the text, the larger its node is.

4.2 Coreference Resolution

A sentence needs not explicitly mention the names of the two people when describing their interaction; the most common alternative is the use of pronouns. Consider the sequences of sentences 'Joseph had been brought down to Egypt … An Egyptian named Potiphar purchased him.' The 'him' clearly refers to Joseph. Whereas the baseline (Section 4.1) misses this relation between 'Joseph' and 'Potiphar', coreference information would enable a link to be established between the two.

With coreference information, recall is expected to improve. However, the accuracy of coreference resolution systems tend to deteriorate as the distance between the pronoun and the mention increases. We therefore only take into account those pronouns within *n* sentences of the mention, where *n* is to be tuned on development data.

4.3 Syntactic and Semantic Information

Even when two names co-occur in a sentence, they do not necessarily signal an interaction. Consider the sentence 'Hamor went to speak with Jacob about Dinah'. The proximity of the names 'Hamor' and 'Dinah' does not imply that the two of them were involved in any interaction. Likewise, despite the co-occurrence of 'Hadad' and 'Masrekah' in the sentence 'When Hadad died, Samlah from Masrekah succeeded him as king', it does not follow that Hadad had been to that location.

This section describes our use of a variety of syntactic and semantic information to address this problem. We leverage part-of-speech and dependency information from a state-of-the-art tagger (Toutanova et al., 2003) and dependency parser (De Marneffe et al., 2006), as well as semantic information from FrameNet (Ruppenhofer et al., 2010).

Person-Person Edges: We derived rules from our development data to filter out invalid edges obtained from the baseline.

Implicit descriptions. As described in Section 3.2, these descriptions involve social interactions, typically actions (e.g., 'kiss') performed by one or both of the people concerned (e.g., 'Jacob kissed Rachel and began to weep'). Therefore, to determine whether the two people are involved in a social interaction, we first check whether the two named entities were marked in the dependency tree either as a subject-object pair of a verb, or as a pair connected by a coordinating conjunction (e.g., 'and'), serving as a subject or object.

Furthermore, the verb must belong to a frame in FrameNet that is deemed to indicate social interactions. To be included in this set of frames, the frame must contain at least one word that is annotated as indicating an interaction in the development data (see Section 3.2). There are 316 selected frames, such as `request` and `cause harm`. During evaluation, the verb must belong to one of these frames in order to be counted towards a person-person edge. This procedure excludes frames such as `perception experience`, thereby successfully blocking such verbs as 'overhear' and 'see', which do not require participation from both parties, and thus do not contribute to a person-person edge.

Explicit descriptions. Personal relationships are usually explicitly realized (e.g., 'son'). They could be stated directly, like in the sentence 'The *sons* of Midian were Ephah, Epher, Hanoch, Abida, and Eldaah.' They could also be mentioned in passing, as in the sentence 'But Jacob did not send Joseph's *brother* Benjamin with his brothers.' In both cases, the relationship word and the relevant personal names are related in predictable dependency structure patterns.

To detect these explicit descriptions, we obtained the list of words that fall under the frames `kinship` or `personal relationship` in FrameNet. If the dependency tree of the sentence contains two or more personal names, both linked to one of these words, then an edge was drawn between the two corresponding person-nodes in the social network.

*Limitations***.** We do not yet handle personal mentions that require compositional analysis. For example, in the sentence 'Sarah noticed the son of Hagar mocking', it was the son of Hagar, instead of Hagar herself, who was being referred to. In general, a noun phrase of the form "X of Y", where Y is a personal name and X is a noun belonging to the `kinship` or `personal relationship` frame, usually refers not to Y but to someone else. Such personal names are therefore ignored. The same policy applies to geographical names requiring compositional analysis, such as ‘south of <location>’.

Person-Location Edges: A geographical name indicates the location at which a scene takes place. Once the scene is established, the location may not appear again in the text. For example, the sentence ‘Joseph had been brought down to Egypt’ is followed by the sentence ‘An Egyptian named Potiphar purchased him.’ It is clear that both Joseph and Potiphar were physically at Egypt. Whenever a geographical name does not appear with the relevant personal names in the same sentence, the baseline would fail to infer the person-location edges.

In order to improve the recall of these edges, whenever a geographical name is detected in a sentence, it is set as the ‘current-location’. Any person mentioned in subsequent sentences is assumed to be present at that location, and an edge is drawn between the current location and that person. This continues until the next geographical name is detected, and the current-location updated.

A naive application of this strategy would, however, result in spurious associations between locations and personal names, since some locations are mentioned only in passing. For example, the location ‘Egypt’ in the sentence ‘They finished eating the grain they had brought from Egypt’ is only used to describe a property of the grain, rather than indicating a change of scene. The constituent in which the geographical name is located can help flag these cases; in particular, prepositions and relative clauses are good indicators.

Prepositions. If a geographical name is preceded by the preposition ‘from’, the location is often used for describing the origin of a person or an object, rather than a change of scene. Such geographical names, therefore, were not set as current locations but were only matched with the personal names that appeared in the same sentence.

Relative clauses. Relative clauses can also be used to determine whether the geographical name should be set as the current location. Geographical names within relative clauses are mainly used to describe a person or the position of another location, and should not be considered as a change of scene[6].

There is one exception. If the head of the relative clause is linked to a personal name, then any geographical names found within the clause are matched to that person[7].

Motion verbs. There is a third phenomenon, where the ‘current-location’ becomes unknown. Motion verbs, such as ‘go out’ and ‘travel’, suggest a change of scene, but the destination is not always specified. When a motion verb is not accompanied with a new geographical name (e.g., ‘he left’), the current location is reset and becomes ‘unknown’; subsequent sentences are not associated with a scene until the next current-location is found. All verbs in the `motion` frame in FrameNet are considered to have this property.

4.4 Baseline using Machine Learning

As a second baseline, we cast the problem of network extraction as a classification task. Two maximum-entropy classifiers (Bird et al., 2009) were trained. One determines whether to connect two person-nodes in the network; the other decides whether to connect a person to a location. As shown in Table 1, most of their features replicate those in the proposed algorithm (Section 4.3), with an additional feature for POS information that further improved performance.

Person-Person Edges	**Person-Location Edges**
Verbs connected to both names in tree	Prepositions heading the names
Presence of words in FrameNet indicating a personal relationship	Whether the name is designated as the current location
Dependency between name and its head	Whether the names are found within relative clauses
Distance between names in sentence	POS of names and surrounding words
POS of names and surrounding words	

Table 1: Features of the classifier for person-person edges and those for person-location edges.

5 Data

The first five books in the Hebrew Bible, or Old Testament, were used for evaluation. We used an

[6] To isolate such clauses, we made use of the dependency tree, which used the label `rcmod` to link the head of a relative clause to the main sentence.

[7] E.g., ‘Hadad’ should be linked to ‘Moab’ in the sentence “[Hadad] … who defeated the Midianites in [Moab], reigned in his place.”

online, open-source English translation known as the New English Translation (NET, 2006). This corpus was chosen for two reasons. First, these five books, also known as the Pentateuch, contain a variety of writing style, from the mostly first-person account in Deuteronomy, and the commands and imperatives in Leviticus, to the narratives in the rest. It is a challenging corpus that can reveal the extent to which our algorithm can generalize. Second, as a well-read corpus, there are a lot of existing resources to enrich our evaluations. For example, we made use of previous published biographies (see Section 6.3).

In the proposed approach, the first book in the Pentateuch, *Genesis*, was used as development set, and the four remaining books, *Exodus*, *Leviticus*, *Numbers* and *Deuteronomy*, as test set. In the machine learning approach, for each book in the test set, a classifier is trained on the rest of the Pentateuch. The network graphs of all five books were drawn manually by annotating sentences according to the criteria set out in Section 3.2. Statistics of the test data are presented in Table 2.

	Exod.	Lev.	Num.	Deut.
# words	31257	23876	30465	25610
# sentences	1371	866	1452	1022
# P-nodes	9	7	27	14
# P-P edges	13	4	32	18
# L-nodes	30	7	116	76
# P-L edges	46	2	114	67

Table 2: Size of our test data. Statistics on the social network graphs include only those characters used in our evaluation, i.e. those mentioned ten times or more. ‘P’ stands for ‘person’, and ‘L’ for ‘location’.

6 Evaluation

This section describes some data processing steps (Section 6.1), then reports experimental results (Section 6.2), and ends with an evaluation from a different perspective, using biographies written by humans (Section 6.3).

6.1 Data Preparation

We extracted named entities from our corpus using the Stanford NER tagger (Finkel et al., 2005). On the test set, for identifying the person-nodes, the tagger yielded 82.1% precision and 71.1% recall; for identifying the location nodes, it yielded only 37.8% precision and 56.7% recall. As for coreference resolution, we made use of the Stanford Deterministic Coreference Resolution System (Lee et al., 2011; Raghunathan et al., 2010).

Since it is common for characters to be referred to with multiple names, we employed the name clustering method in Elson et al. (2010), matching the named entities with their variations.

6.2 Results

We first analyze the results for person-person edges and person-location edges, using named entities extracted manually (gold named entities). We then report the effects of using automatic named entity recognition. In all evaluations, we considered only the major characters, defined as those mentioned at least ten times in the corpus.

Algorithm	Exod.	Lev.	Num.	Deut
Baseline	P: 0.43	0.40	0.35	0.53
	R: 1.00	1.00	0.97	0.89
	F: 0.60	**0.57**	**0.51**	**0.67**
Classifier	P: 0.65	0.50	0.69	0.64
	R: 0.85	1.00	0.69	0.50
	F: 0.73	**0.67**	**0.69**	**0.56**
Proposed	P: 0.59	0.67	0.64	0.58
	R: 1.00	1.00	0.78	0.61
	F: 0.74	**0.80**	**0.70**	**0.59**

Table 3: Precision (P), recall (R), and F-measure (F) of person-person edges in the automatically generated networks. Gold named entities are used.

Person-Person Edges: Experimental results are shown in Table 3. Overall, the proposed approach yielded an average F-measure of 0.71, an improvement [8] over both the baseline and the classifier. Whereas the baseline favors recall, and the classifier favors precision, the proposed approach strikes a balance between the two. It has the added benefit of requiring less training data than the classifier.

In all books except Deuteronomy, gains over the baseline came from improvement in the precision. In particular, the dependency

[8] The improvement is statistically significant for the first three books against both the baseline (p<0.0001 by McNemar’s test) and the classifier (p<0.02).

information was able to discount name pairs that simply happen to be in the same sentence but do not concern one another. Furthermore, the filtering steps using FrameNet detected those that, despite being closely related grammatically (e.g., subject-object), do not involve interactions. Deuteronomy, which consists of mostly first-person, direct speech, proved to be more challenging.

Most mistakes in other books were caused by inaccuracy in coreference resolution, especially plural pronouns. As a typical case, the word 'they' in a sentence[9] refers to two characters, Nadab and Abihu, mentioned earlier as Aaron's sons. The coreference resolution unfortunately linked the word to Aaron himself, resulting in an extra edge and two missed edges.

Another source of error was inaccuracy in dependency parsing, particularly for explicit descriptions in sentences with multiple names. For example, in the sentence 'Now these are the names of the men who are to help you: from Reuben, Elizur son of Shedeur', the word 'son' was wrongly linked to Reuben, instead of Elizur.

Despite the improvement in precision, our proposed algorithm still extracted some extra edges because of ambiguity in meaning. Consider the sentence 'Then Miriam and Aaron spoke against Moses because of the Cushite woman he had married'. Since the verb 'speak' suggests an interaction, our algorithm reckoned this as a social relation. According to our definition, however, a social relationship is recorded only if both parties are aware of the interaction, and so this edge was not marked by the annotator.

Algorithm	Exod.	Lev.	Num.	Deut.
Baseline	P: 0.48	0.15	0.37	0.22
	R: 0.54	1.00	0.55	0.39
	F: 0.51	**0.27**	**0.44**	**0.28**
Classifier	P: 0.50	0.50	0.40	0.38
	R: 0.46	1.00	0.24	0.30
	F: 0.48	**0.67**	**0.30**	**0.33**
Proposed	P: 0.50	0.29	0.46	0.31
	R: 0.61	1.00	0.46	0.39
	F: 0.55	**0.44**	**0.46**	**0.34**

Table 4: Precision (P), recall (R) and F-measure (F) of person-location edges in the automatically generated networks. Gold named entities are used.

Person-Location Edges: Experimental results for person-location edges are shown in Table 4. Our proposed algorithm improved[10] the average F-measure over both the baseline and the classifier. Similar to person-person edges, most gains were due to improved precision, contributed by the filtering performed with prepositions and relative clauses (Section 4.3).

Mistakes in the coreference resolution system, again, were responsible for many missed relations. For example, the sentence 'They were the men who were speaking to Pharaoh king of Egypt' was preceded by a list of more names, all of which should be linked to 'Egypt'. Also, in a number of cases, the personal names appeared before the location. Our strategy of maintaining the current-location failed to connect these names to the location.

Automatic named entity recognition: If named entities in the corpus are automatically extracted, mistakes in NER would trickle down to the social network. Unsurprisingly, both precision and recall deteriorated in most books, resulting in an average precision of 0.55, an average recall of 0.32 and an average F-measure of 0.40 for person-person edges, an average precision of 0.07, an average recall of 0.20 and an average F-measure of 0.09 for person-location edges.

6.3 Comparison with Biographies

For many well-known works of literature, including our evaluation corpus, there already exist human analyses of the characters and their inter-relationships, in the form of biographies. To provide a different angle of evaluation, we measure how these biographies differ from the kind of social networks constructed by our algorithm, using the book *Who's Who in the Old Testament* (Comay, 2001), which provides sketches of the lives of a number of major characters.

Out of these biographies, we constructed social networks by first inserting a node for each character that appears in the Pentateuch. We then scanned for personal and geographical names in the biography, and added edges between that node

[9] In 'So fire went out from the presence of the Lord and consumed them so that they died before the Lord'.

[10] The improvement is statistically significant against the baseline for the book of Exodus ($p < 0.01$ by McNemar's test), and against the classifier for Deuteronomy ($p < 0.002$).

and the corresponding nodes representing those names.

The social networks constructed from these biographies are compared to our manually annotated ones. They yielded an average precision of 0.19, an average recall of 0.75 and an average F-measure of 0.29 for person-person edges; an average precision of 0.10, an average recall of 0.30 and an average F-measure of 0.14 for person-location edges. Both the precision and recall are substantially lower than the proposed algorithm.

These results must be qualified in two respects. First, although only the biographies for those characters that appear in the particular book under evaluation were considered, they still contain information on events that occurred outside of the book. Further, the biography-based networks were constructed with expert knowledge, and may include, therefore, social relations that are implied but without textual evidence. These mismatches with the gold networks contributed to a lower precision.

Second, certain social interactions may be deemed by the author as insignificant and therefore omitted; in contrast, no such judgment was made in our annotations. This led to a lower recall.

7 Conclusion and Future Work

We have described and evaluated an algorithm that automatically infers social networks from literary texts. The algorithm outperforms a co-occurrence baseline as well as a statistical classifier. A significant novelty of these networks is that they encode not only people and their relations, but also the locations at which they are active, and the sentences that attest to these claims. Readers can browse a higher-level view of the relationships among characters, and easily refer to the relevant sentences.

We plan to build on this work in several directions. First, we would like to improve the precision and recall of the automatically generated networks, by borrowing more techniques from relevant fields in natural language processing. Second, we intend to generalize our algorithm to other languages, so as to generate networks for international literary works. Third, it would be useful to further characterize the nature of the edges, such as whether two people are “friends” or “foes” (van de Camp and van den Bosch, 2010), and the kind of activities that a person is engaged at a location.

Acknowledgments

This research project was partially supported by a CityU Start-up Grant for New Staff.

References

Apoorv Agarwal and Owen Rambow. 2010. Automatic detection and classification of social events. *Proc. EMNLP*.

Apoorv Agarwal, Owen Rambow, and Rebecca J. Passonneau. 2010. Annotation Scheme for Social Network Extraction from Text. *Proc. ACL*.

Apoorv Agarwal, Owen Rambow, and Rebecca J. Passonneau. 2010. Annotation scheme for social network extraction from text. *Proc. Fourth Linguistic Annotation Workshop*.

NET 2006. *The Net Bible*. Biblical Studies Press.

Marcus Bingenheimer, Jen-Jou Hung, and Simon Wiles. 2009. Markup meets GIS – Visualizing the ‘Biographies of Eminent Buddhist Monks’. *Proc. 13th International Conference on Information Visualisation*.

Steven Bird, Edward Loper and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Nancy Chinchor, Erica Brown, Lisa Ferro, and Patty Robinson. 1999. *Named Entity Recognition Task Definition*. Technical Report, MITRE Corporation and SAIC.

Joan Comay. 2001. Who’s Who in the Old Testament (Who’s Who (Routledge)). Routledge.

Jana Diesner, Terrill Frantz, and Kathleen Carley. 2005. Communication Networks from the Enron Email Corpus: It’s Always about the People, Enron is no Different. *Computational and Mathematical Organization Theory* 11(3).

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program: Tasks, Data, and Evaluation. *Proc. LREC*.

Gregory Crane. 2006. What Do You Do with a Million Books? *D-Lib Magazine* 12(3).

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. *Proc. LREC*.

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. 2010. Extracting social networks from literary fiction. *Proc. ACL*.

Jenny R. Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. *Proc. ACL*.

Stephan Greene, Gary Marchionini, Catherine Plaisant, and Ben Shneiderman. 2000. Previews and Oerviews in Digital Libraries: Designing Surrogates to Support Visual Information Seeking. *Journal of the American Society for Information Science* 51(4):380—393.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford’s Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. *Proc. CoNLL*.

Inderjeet Mani and George Wilson. 2000. Robust Temporal Processing of News. *Proc. ACL*.

Daniel Marcu. 1998. The Rhetorical Parsing of Natural Language Texts. *Proc. ACL*.

Franco Moretti. 1999. *Atlas of the European Novel 1800-1900*. Verso.

Paul Mutton. 2004. Inferring and Visualizing Social Networks on Internet Relay Chat. *Proc. 8th International Conference on Information Visualization*.

NET 2006. *The Net Bible*. Biblical Studies Press.

James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Saurí. 2005. Temporal and Event Information in Natural Language Text. Language Resources and Evaluation 39:123---164.

Josef Ruppenhofer, Michael Ellsworth, Miriam Petruck, Christopher Johnson, and Jan Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. http://framenet.icsi.berkeley.edu

Jeff Rydberg-Cox. 2011. Social Networks and the Language of Greek Tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1(3).

David A. Smith. 2002. Detecting and Browsing Events in Unstructured Text. *Proc. SIGIR*.

David A. Smith and Gregory Crane. 2001. Disambiguating Geographical Names in a Historical Digital Library. *Proc. ECDL*.

Michael Speriosu, Travis Brown, Taaesun Moon, Jason Baldridge, and Katrin Erk. 2010. Connecting Language and Geography with Region-Topic Models. *Proc. Workshop on Computational Models of Spatial Language Interpretation (COSLI)*.

James Stiller, Daniel Nettle, and Robin I. M. Dunbar. 2003. The Small World of Shakespeare’s Plays. *Human Nature* 14(4):397---408.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proc. NAACL-HLT*.

Matje van de Camp and Antal van den Bosch. 2011. A Link to the Past: Constructing Historical Social Networks. *Proc. Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*.