Towards an Automated Approach to Extract and Compare Fictional Networks: An Initial Evaluation

Marcello Trovati and
James Brady
School of Computing and Mathematics
University of Derby
Kedleston Road, Derby, UK
Email: M.Trovati@derby.ac.uk

***Abstract*—Within the field of Digital Humanities, fictional networks are defined by the social interactions among characters, capturing unique features associated with each different textual source. The ability to automatically extract, analyse, and assess such properties would open new research directions, and lead to relevant and important results. However, these are challenging tasks which require a multi-disciplinary approach due to their inherent complexity.**

In this paper, we will describe the initial implementation of a system to address the above. This is part of an ongoing research investigation aiming to define and implement a similarity measure to compare specific properties associated with fictional networks extracted from across different genres and sources, as well as from different historical periods. The preliminary evaluation clearly shows the potential of our approach, as well as the applications to a variety of multi-disciplinary research fields.

I. Introduction

Characters in literature are involved in a sequence of interwoven narratives which create a rich tapestry of social interactions. These are known as *fictional networks*, an emerging multi-disciplinary area of research which has steadily gained momentum especially within the domain of Digital Humanities [8]. In particular, there have been attempts to integrate their analysis with text mining to automate the process of identifying and classifying relevant information from literature [2].

There is an extensive set of text mining tools and techniques which can be applied to this field [1]. However, during our investigation, we have ascertained that a more agile and lightweight method is more likely to produce accurate results, as well as being fully scalable.

In this paper, we will describe some preliminary results which are part of an ongoing and wider line of inquiry to extract, investigate, and assess fictional networks and their topology. In particular, this includes the introduction of a similarity measure to efficiently and accurately compare texts in terms of the social interactions between characters. However, this poses a challenging task as, apart from the extraction of the characters mutually connected by social interactions, it is also necessary to classify the type of such relations, both quantitatively and qualitatively, as well as chronologically. This paper focuses on the extraction of fictional networks, which includes an evaluation carried out by analysing "The Legends Of King Arthur And His Knights" by James Knowles [6]. This produced interesting results, suggesting new research directions, and the potential of our approach.

II. Background

A. Network theory

Network theory is a branch of graph theory with very broad applicability. Mathematically speaking, a network $G = (V, E)$ consists of a set of nodes $V = \{v_i\}_{i=1}^{n}$ and a set of edges $E = \{e_{ij}\}_{i \neq j=0}^{n}$, where $e_{lm} \in E$ if and only if v_l and v_m are connected by an edge. A network is said to be *directed* if its edges are not commutative, or in other words, $e_{ij} \neq e_{ji}$. Otherwise, it is said to be *undirected*.

Despite the simplicity of the above definitions, networks describe and model complex theoretical and real-world applications. These include the study of transport, polymers, economics, particle physics, computer science, sociology, epidemiology, linguistics and more [9]. Another notable example includes the investigation of social networks, where nodes usually correspond to people, and edges represent a social interaction between any two of them. In particular, their main properties are successfully described by network topology, which in turn, can be used to investigate and predict their overall behaviour.

B. Natural Language Processing in Digital Humanities

The clear advantage in using any computerised information extraction from text, is the potentially large set of textual information that can be processed. However, this is one of the most challenging tasks in computer science, due to the highly ambiguous nature of human language. In fact, its interpretation depends on a huge number of factors which can be difficult to fully implement on a machine.

Natural Language Processing (NLP) [7], p. 70, is a branch of computer science which aims to accurately extract, identify and analyse information and semantic properties from text sources. Even though there has been steady and successful progress in addressing the above challenges, NLP research is still very much expanding to provide further state-of-the-art tools to improve accuracy, scalability and flexibility. There is extensive research on the automated extraction of fictional networks from text, which mainly focuses on the statistical applications of the co-occurrence of characters. Broadly speaking, this relates characters if they occur within the same fragments of text (see the bibliographies in [2], and [10] for more details). However, one of the main criticisms is that if two characters occur in the same fragment, it does not necessarily imply that they are linked in a conclusive way. In [2], a more sophisticated

approach is used. In this case, relationships are identified whenever two characters are linked by a direct conversation.
On the other hand, grammar based text extraction relies on a set of rules which identify sentences with a determined structure. The effectiveness of this approach is fully exploited when syntactic properties of a sentence are investigated, by using suitable parsing technology [7], p. 107. In particular, the syntactic roles of the different phrasal components are essential in extracting the relevant information, and they can contribute towards a full understanding of the type of relationship. This is the approach used in [10] where nouns linked by a verb are extracted as triplets. However, there is little attempt to address some important issues, such as collocations and the identification of the relation types, as well as the improvement of of grammar-based methods to provide a more accurate extraction.

III. Description of our Approach

As mentioned above, automatically extracting and analysing a variety of fictional networks from across different genres and from different historical periods, would open new research directions within a multitude of disciplines.
As part of our current investigation, we have been designing a system to address the initial stages of this analysis. Broadly speaking, it has the following feature

- The text analysis is based on the Stanford parser technology, including its Named Entity Recognition (NER), co-reference and anaphora resolution tools [3].
- Two dictionaries were manually constructed. The former contains common fictional characters with their different spellings to ensure they would be identified correctly. The latter includes sets of verbs divided into *hostile* and *friendly* ones, which were manually identified from the data set introduced by Hu and Liu [4]. Table I shows a small selection of such verbs. This is used to identify the type of relations, i.e. friendly vs hostile, between characters.
- Some general collocation issues are addressed, as discussed in Section III-A. This has proved particularly efficient when analysing old English texts. In fact, we noticed that the relation extraction was considerably improved when utilising the algorithm.
- The relations between characters are extracted via grammar and co-occurrence based approaches. These are then categorised according to their type, namely whether they correspond to a *friendly*, *hostile*, or *unknown* relation, as described in Section IV.
- Once the relation extraction and classification have been carried out, a simple aggregation method is used.

In the following sections we will describe the above components of the system in detail.

A. Collocation and Character Identification

In [5], a term detection method is introduced, where the authors discuss some techniques applicable to collocation resolution and in particular character identification. Their approach

TABLE I. A selection of friendly and hostile verbs contained in the dictionary as described in Section III.

Hostile verbs	Friendly verbs
kill	accommodate
hate	admire
injure	aid
harm	approve
attack	approve
annihilate	cherish
asphyxiate	cooperate
assassinate	collaborate
crucify	cuddle
drown	esteem
eradicate	fondle
erase	glorify
execute	idolise
exterminate	kiss
extirpate	love

is based on the use of specific patterns as shown in Table II, which produces accurate results. In this paper, we have

TABLE II. Text patterns for term extraction as in [5].

Tag Pattern	Example
Adj Noun	linear function
Noun Noun	regression coefficients
Adj Adj Noun	Gaussian random variable
Adj Noun Noun	cumulative distribution function
Noun Adj Noun	mean squared error
Noun Noun Noun	class probability function
Noun Prep Noun	degrees of freedom

used a similar approach, yet specifically designed for the identification of fictional characters, in order to specifically address characters' collocations in old English corpora. In fact, the use of text patterns, along with the NER technology offered by the Stanford Parser, has shown to accurately determine the main characters in the fictional literature we have analysed. More specifically, we defined the following cases:

1) `title` + `prep` + (`adj`+) `noun` will be identified as a character if
   - `title` specifies someone's position, e.g. "King", "Sir", "Dame", etc.
   - `prep` is one preposition, such as "of" as well as a determinant, such as "the", or a combination of more than one preposition
   - `adj` is an optional adjective
   - `noun` refers to a geographical area (e.g. Ireland), or people (e.g. Jews), so that "King of Scotland", "King of the Jews", "priest from Ireland" are successfully identified.
2) `title` + `noun` will also be identified as a character if
   - `title` specifies someone's position, as above
   - `noun` refers to a proper noun, so that "King Death", is successfully identified.
3) Groups of words indicating possession, i.e. `adj` + `name_1's` + `adj` + `name_2`, where
   - `name_1` and `name_2` are nouns referring to persons, e.g. "Paul's wife"
   - `adj` is an adjective. So "King's great servant" would be recognised as unique entities.

This approach has been shown to produce good results even though this has only been tested on Old English texts. We are planning to expand and improve this method to incorporate the relevant linguistic and algorithmic properties.

B. Relation Extraction

At this stage, the relation extraction was based on both grammar and co-occurrence based approaches.
Grammar based methods refer to the model describing the process of generating structured sentences within human language [7]. This structure is described via the part-of-speech (POS) tagging which labels words according to their definition and their relationships with related words in a phrase. A useful feature provided by the majority of parsers is the capability of identifying the "tree structure of sentences". In fact, the syntactic role of words naturally create a hierarchical tree structure which can be used to identify fragments which exhibit a specific structure. This is usually carried out via *text patterns*. As the name suggests, they aim to capture sentences that contain particular syntactic patterns. There is a huge range of possible text patterns to address most of the requirements of data extraction. However, defining effective text patterns is a lengthy process where often, a trial and error approach is used. In this paper, we focused on text patterns with a subject-verb-object structure, or triplets [10], which are often described as

```
<NP1, verb, NP2>
```

where `NP1` and `NP2` are the noun phrases, and `verb` is the linking verb. As we are interested in extracting characters' names contained in the noun phrases, these need to be further analysed to isolate the correct names. There are clearly a variety of limitations with this simple type of extraction since only sentences with this particular structure can be identified, which can potentially affect the overall extraction output.
However, natural languages are ambiguous by nature since words may refer to more than a single meaning depending on the context. Therefore, POS tagging is typically a hard problem whose accuracy depends on the complexity of the syntactic and semantic structure of a text fragment. In fact, disambiguation techniques are at the heart of NLP research [7]. In particular, the efficiency of the above text pattern depends on the type of text analysed, as well as on its semantic and lexical properties. As a consequence, couples consisting of co-occurring characters within text fragments have also been analysed so that the overall extraction could be improved. However, in order to minimise any potential inaccuracy, we defined the following algorithm.

Relation Extraction Algorithm.

Let W_P be the set of words in the text recognised as persons, W_G be the set of couples (w_i, w_j) such that w_i and w_j are characters extracted via grammar based rules, and W_F be the output list of the algorithm.

1: Let $W_F := W_G$. In other words, W_F is first defined to have the same elements as W_G.
2: **for all** (w_i, w_j), such that w_i and $w_j \in W_P$, $i \neq j$ **do**
3: **if** $(w_i, w_j) \in W_F$ **then**
4: discard the couple.
5: **else**
6: **if** there exists $w_k \in W_P$, such that $(w_i, w_k) \in W_F$, and $(w_k, w_j) \in W_F$, $(w_i, w_j) \notin W_F$ **then**
7: add (w_i, w_j) to W_F.
8: **end if**
9: **end if**
10: **end for**
11: **return** W_F

Figure 1 depicts the general process described in this section.

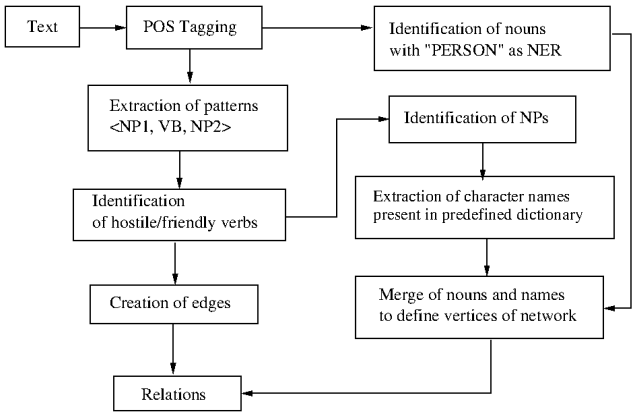

Fig. 1. An overview of the process described in Section III-B.

Note that at this stage, all the couples are commutative, that is $(w_i, w_j) = (w_j, w_i)$. In other words, fictional networks are undirected. In fact, determining the direction of a relation via text extraction, implies a much deeper and complex linguistic analysis, which goes beyond the scope of this paper.
Finally, the relations extracted consist of triples of the form

```
(char_1, char_2, type)
```

where the first two entries correspond to the pair of fictional characters linked by the relation, and the last one specifies its type, which depends on the nature of the linking verb as discussed above, or whether they have been isolated via the co-occurrence algorithm, as described in Section IV.
Furthermore, in the case of sentences where more characters are linked by a verb, we assume that they are mutually linked to each other. We acknowledge that this approach is reductive since it does not distinguish between conjunctions and disjunctions, and we are aiming to address this further in future research.

IV. Identification of the Relation Types

Another aim of our research is to fully identify and assess the types of relationships between characters. At this stage, we have made the following assumptions:

- *Dynamical nature of types.* Any relation can change and evolve into a different type over time.
- *Non-commutativity of types.* A relation between two individual is, in general, non-commutative. Therefore, it would be intuitive to consider the network generated

by the relations classified according to their types, as a directed one. However, this would create a variety of critical challenges as mentioned above. We aim to expand this topic in future research.

As discussed above, the analysis of each fragment is carried out producing sets of triples corresponding to the two characters linked by a relationship, and its type. They can be

- *Friendly* or *hostile*, depending on the nature of the linking verb, and
- *Unknown*, if the relation fails to be categorised.

All the triples are then aggregated according to the following algorithm.

Type Aggregation Algorithm.

1: Remove any relation that has not been recognised as either friendly or hostile (i.e. unknown), and group the remaining ones into a set L.
2: Let $l_1, \ldots, l_j$ be its (disjoint) subgroups whose elements are the triples referring to the same two characters. Clearly, $\bigcup_{i=1}^{j} l_i = L$, and $\sum_{i=1}^{j} |l_i| = |L|$, where $|\cdot|$ refers to the cardinality of a set.
3: **for all** l_i, $1 \leq i \leq j$ **do**
4: assign the most frequently occurring type to the relation between the two characters.
5: **end for**

For example, for a relation between two characters that appears with a higher number of friendly types than hostile ones, then assume it is of friendly type. Note that, a full type aggregation, would also require a chronological sorting to assess the evolving of each relation. However, this is typically hard as implicit semantic information is needed to be fully analysed [12]. In future research, we are aiming to carry out a thorough investigation of this particular issue.

V. Evaluation

A formal evaluation was carried out on the first volume of the James Knowles' "The Legends Of King Arthur And His Knights" [6]. Once the automated extraction had been carried out, it was compared with the manually extracted relations from the same text for validation purposes. Note that the manual extraction output consisted of couples of characters linked by a relation and its type, to make it consistent with the output of our system. This produced a precision of 89% and the recall of 68%.
Furthermore, 55% of the (automatically) correctly extracted relations also had a correct type. Interestingly, we also evaluated a different basic aggregation method based on just considering the type of the *last* instance of a relation. This follows the assumption that this is likely to be the most recent (and so a definitive) type of relation. In such case, only 41% of the (automatically) correctly extracted relations had the correct type, suggesting that it might not be a feasible approach. As discussed in Sections III and IV, we only used a relatively limited set of dictionaries and text patterns. Therefore, the recall is understandably affected. On the other hand, the efficiency of the implementation of our system is shown by the high value of precision. As far as the relation types are concerned, considering the nature of the linking verb, only captures a limited set of possible types, which is reflected by the relatively low level of correct type extraction. In fact, most of the types wrongly classified were identified as *unknown*.
The evaluation of the topology of the fictional network was also carried out as shown in Table III. In particular, the network automatically extracted has an increased average degree and clustering coefficient with respect to the manually extracted one, clearly suggesting that the extracted characters cluster together more than they should, with an average path length reduced by 9%.

TABLE III. The evaluation of the topology of the fictional network as described in Section V

	Original	**Extracted**
No. vertices	283	298
No. edges	1238	1417
Average Degree	8.53	9.22
Average Path Length	1.98	2.07

VI. Discussion

Although the idea of automatically mine literature for relevant information is certainly not new, our approach is specifically designed to describe some crucial mathematical issues related to network analysis [9], [11]. As discussed above, our main motivation is based on the fact that the full description of network topology requires a variety of information, including the full identification of nodes, as well as the classification of the types of the edges between them. This would allow the identification of a variety of important aspects, including

- Clustering properties,
- General dynamical properties,
- Identification of general topological features, such as scale-free and small-world behaviour, applicable to both the whole corresponding network, and its sub-components [11].

The method introduced in this paper is an important part of the development of an agile method to assess similarity between fictional network, based on the dynamical and topological properties of fictional networks. Two different characters, not necessarily from the same novel, can be compared via the features exhibited by the corresponding networks, and their sub-components. However, this can become easily demanding from a computational point of view due to the complexity, and size of fictional networks [2], [8]. The preliminary design and implementation of our approach show promising results in terms of scalability, robustness, and computational efficiency. This will be the focus of a future paper submission.

VII. Conclusion

In this paper, we have introduced and discussed the initial steps of a system which automatically extracts, assesses and

analyses fictional networks extracted from text. Our preliminary evaluation shows promising results even though some components of the system still have limited capabilities. This clearly shows the potential of this approach and in order to further improve its performance, we aim to carry out the following:

- Improve and extend the text analysis to extract a wider set of relations.
- Fully address the identification of the type of relations, including their direction, to allow a more accurate assessment.
- Carry out a full investigation of the relation aggregation to incorporate all the relevant linguistic and dynamical properties.

As mentioned above, the system introduced in this paper is part of a wider line of inquiry focusing on the definition, identification, and evaluation of a similarity measure to compare fictional networks. In fact, being able to identify characters that exhibit the same topological properties, would enable the investigation of similarities between fictional networks associated with different text corpora, and open a multitude of new research directions and outcomes.

REFERENCES

[1] H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan, *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*, Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), 2002.

[2] D. K. Elson, N. Dames, and K. McKeown, *Extracting social networks from literary fiction*, Jan Hajic, Sandra Carberry, and Stephen Clark, editors, ACL, pages 138-147, 2010.

[3] J. R. Finkel, T. Grenager, and C. Manning, *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling*, roceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370, 2005.

[4] M. Hu and B. Liu, *Mining and Summarizing Customer Reviews*, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, pages 168177, New York, NY, USA. ACM, 2004.

[5] J. S. Justeson and S. M. Katz, *Technical Terminology: Some Linguistic Properties and an Algorithm for Identication in Text*, Natural Language Engineering, 1(1):927, 1995.

[6] J. Knowles, *The Legends Of King Arthur And His Knights* Retrieved from the Gutenberg Project: http://www.gutenberg.org/files/12753/12753-h/12753-h.htm, 1862.

[7] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts, 1999.

[8] P. Mac Carron and R. Kenna, *Universal Properties of Mythological Networks*, CoRR,abs/1205.4324, 1999.

[9] M. E. J Newman, *The Structure and Function of Complex Networks*, SIAM Review, 2003.

[10] S. Sudhahar and N. Cristianini, *Automated Analysis of Narrative Content for Digital Humanities*, International Journal of Advanced Computer Science, 3(9):440447, 2013.

[11] M. Trovati, N. Bessis, A. Huber, A. Zelenkauskaite, and E. Asimakopoulou, *Extraction, Identification and Ranking of Network Structures from Data Sets*, to appear in the Proceedings of CISYS, 2014.

[12] N. UzZaman and J. F. Allen, *Trips and Trios System for Tempeval-2: Extracting Temporal Information from Text*, Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, pages 276–283, Stroudsburg, PA, USA. Association for Computational Linguistics, 2010.