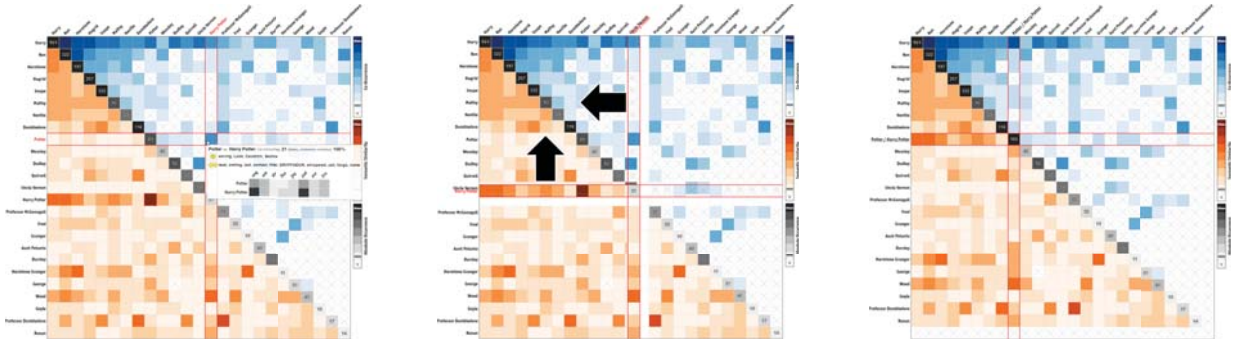
Erschienen in: EuroVis 2017 Eurographics / IEEE VGTC Conference on Visualization 2017 / Heer, Jeffrey et al. (Hrsg.). - (Computer Graphics Forum ; 36,3). - S. 189-200
https://dx.doi.org/10.1111/cgf.13179

Interactive Ambiguity Resolution of Named Entities in Fictional Literature

Florian Stoffel, Wolfgang Jentner, Michael Behrisch, Johannes Fuchs and Daniel Keim

University of Konstanz, Germany



(a) *Hovering a cell highlights the respective row (Potter), column (Harry Potter), and their labels. Dark cells that spot out indicate possible entities to merge.*

(b) *During the animation, the right column and lower row (both highlighted) are moved to the left and up respectively until they completely overlap the column and row of the second entity.*

(c) *The matrix after the merge process: Potter and Harry Potter are now merged and their co-occurrences and similarities are updated accordingly, the corresponding row and column are highlighted.*

Figure 1: *The process of merging entities in* AMBIGUITYMATRIX *from left to right. On the left, the initial state of the visualization is shown. Users can explore the visualization, as well as the displayed data by tooltips. The middle image shows the process of merging, the prior selected row and column are moved on to the top left, as the arrows indicate. The right image depicts the state after the merge process. The visualizations are created with data from Harry Potter and the Sorcerer's Stone by J. K. Rowling.*

Abstract

Named entity recognition (NER) denotes the task to detect entities and their corresponding classes, such as person *or* location*, in unstructured text data. For most applications, state of the art NER software is producing reasonable results. However, as a consequence of the methodological limitations and the well-known pitfalls when analyzing natural language data, the NER results are likely to contain ambiguities. In this paper, we present an interactive NER ambiguity resolution technique, which enables users to create (post-processing) rules for named entity recognition data based on the content and entity context of the analyzed documents. We specifically address the problem that in use-cases where ambiguities are problematic, such as the attribution of fictional characters with traits, it is often unfeasible to train models on custom data to improve state of the art NER software. We derive an iterative process model for improving NER results, show an interactive NER ambiguity resolution prototype, illustrate our approach with contemporary literature, and discuss our work and future research.*

1. Introduction

Named entity recognition (NER) techniques are used to identify passages of text that are likely to refer to entities and label them with categories such as *person*, *place*, *company* and other categories [NS]. Current state of the art techniques are based on manually annotated text corpora, that are automatically analyzed and transferred in a corresponding model of language use, grammar, and other properties of the annotated text. These models can be used to find and classify entities in previously unknown documents. State of the art NER techniques are provided with off the shelf models created from

huge amounts of annotated documents, such as news corpora or similar document collections. For many use-cases, the standard models perform well and are well-suited for the integration in different applications, such as the analysis of customer reviews [LZ12], or the automated extraction of locations or characters in literature [FG15]. Whenever these techniques and models are applied to documents that contain uncommon or underrepresented linguistic characteristics compared to the training data, the performance of NER systems degrade, which in consequences makes errors more likely to happen. To cope with that problem, state of the art NER packages provide facilities to create custom models from a collection of documents. Given that for the application at hand enough data is available, text annotation requires domain experts to annotate the data in adequate amount and quality, which is a time-consuming and difficult process [RC12].

Charlie Weasley	George Weasley	
Fred Weasley	Percy Weasley	Ron Weasley
Ronald Weasley	Ginny Weasley	Weasleys
Harry Potter	H. Potter	Potter
The Potters	James Potter	

Table 1: *A selection of entities classified as* person *that have at least one word in common. Extracted from* Harry Potter and the Sorcerer's Stone *by J. K. Rowling with state of the art NER software.*

In this work, we introduce an interactive, visualization-based approach to improve the performance of NER software with respect to ambiguities, e.g. the multiple occurrences of the same entity caused by different attributions and references, as illustrated in Table 1. With the help of our technique, AMBIGUITYMATRIX, such ambiguities can be identified through model visualization and interactions so that further annotation of the data and the corresponding time-consuming training phase is not required. The visualization makes use of two interlinked triangular matrix plots representing the results of a specific NER model/technique as the row/column appearance, as they can be seen in Figure 2. The color-coded cells in the upper triangular display the pairwise co-occurrence of entities, while the cells in the lower triangular matrix show the semantic similarity of the entity surrounding based on word embeddings [BDVJ03, CW08]. Both views guide the analyst to possible entity ambiguities, indicated by co-occurrences (upper half) or semantic similarity of the character surroundings (lower half). Entities with high co-occurrence and/or high semantic similarity, that we assume are connected to ambiguities of the NER, stand out by a dark color in both matrix triangles. Less obvious, potential ambiguities follow the visual pattern of a single, outstanding cell in one of the two views. By merging two rows/columns, the user can derive disambiguation rules, e.g. the merger of two entities, which can be used to improve NER models in a post-processing stage, or by feeding back the interaction and corresponding data to a model-refinement or recreation phase. We focus on referential ambiguities as their detection and resolution requires a high level of text understanding, which is currently not feasible for natural language processing techniques. In contrast, our approach relies on text understanding and world-knowledge of the user, and can be applied without requiring machine readable, externalized knowledge, for example in form of knowledge bases. For illustration purposes, we focus in this work on fictional literature to account for a great variety of writing styles, language use, and grammar that is typically not contained in training data utilized to create off the shelf NER models. We concentrate on characters (named entities of type *person*) in fictional literature, which is subject to real-world applications as well as current research [MBK14, VJPR15] in the NER area.

In this work, we claim the following contributions: 1) A visualization-based approach for linking of named entities based on referential ambiguity. 2) A novel combination of state of the art in visualization and natural language processing that assists human analysts in the entity linking process. 3) An interactive interface that combines content and context information to support interactive rule building from ambiguous data sources.

The remainder of this paper is structured as follows: In the next section, we introduce related work. Section 3, presents the core of our work containing a process model, details of the employed NER methods, and the interactive visualization technique which we call AMBIGUITYMATRIX. Section 4 showcases our approach on different books and presents the corresponding visualizations. Section 6 contains a discussion and gives future perspectives before we conclude the paper in Section 7.

2. Related Work

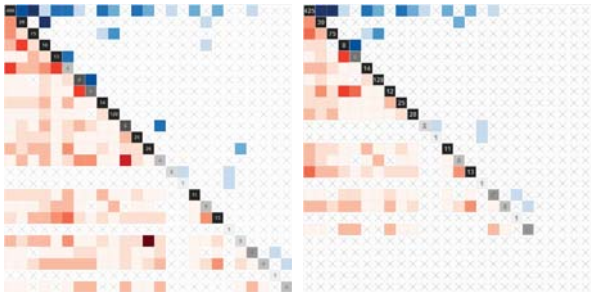
Our work is residing in three different areas: named entity detection, entity or record linking, and entity relationship visualization. In the following, we give an overview of related work in those areas.

2.1. Named Entity Detection

Named entity detection is a part of the techniques collectively called NERC (Named Entity Recognition and Classification) [SR09], which summarizes techniques that identify entities and classify them into categories, e.g. *person*, *organization*, or *location*, in an unstructured source of text. Techniques for entity detection are manifold and utilize various methods. In the following, we outline the major techniques and give some insights into their construction.

Lexicon- or dictionary-based systems come pre-equipped with lists of frequent names or locations, that are used for keyword searches, as well as further adaptions based on statistical techniques [BSAG98, BON03, BP06]. *Statistical* approaches utilize non-specific sources of text or word information, such as large document collections, in order to compute relationship statistics applied in the actual entity detection process [SS04, dSKL04]. These techniques do not require a labeled training set and are independent of the document language, as they purely rely on statistical properties of the documents. Recently, *machine-learning* techniques have gained popularity in the field [BMSW97, BMS00, DC08]. They perform well, can be utilized even without a learning process by distributing the appropriate models, and have are highly adaptability to the processed data. Naturally, a drawback of these methods is the dependence on the model that has been trained from labeled documents. A third family of techniques, based on *handcrafted rules or heuristics*, that can incorporate and employ different properties of the analyzed text documents [Rau91, CM99, DNOO08]. These systems have advantages if there is a consistent structure among the documents to

analyze, that can, for example, be induced by the domain of the document contents or specific writing styles, as these can be exactly defined and modeled in the rule set. Therefore, rule-based NER software is expected to reach high precision, while the recall depends on the variety and flexibility of the rule-set. Last, a number of techniques combine different aspects from the aforementioned named entity detection techniques in order to benefit from strengths of the different techniques and methodologies [SKKS01, NC14]. A comprehensive overview of different methodologies and approaches can be found in [NS, Sar08].



(a) *Initial State* **(b)** *Final State*

Figure 2: AMBIGUITYMATRIX *at the example of George Orwell's 1984, generated by OpenNLP Name Finder. The initial sparse matrix got even sparser, most changes are visible in the upper left half depicting the semantic similarity measure. For visual clarity, we removed the entity labels.*

2.2. Entity/Record Linking

The task of merging different entities to a single one is also known as entity linking, which is nowadays part of many natural language processing and NER systems.

Bunescu and Pasca presented a technique that uses support vector machines to learn rankings from categories and names using data from Wikipedia [BP06]. Similarly, Mihalcea and Csomai exploit the Wikipedia hyperlink structure to associate documents with Wikipedia articles, which can also be used to process and link named entities [MC07]. Other data sources, namely DBpedia and YAGO, are used by Hoffart et al. to compute disambiguation of entities, which could also be used to link entities [HYB*11]. Large scale knowledge-bases are exploited by Lin et al., where the authors propose corpus level features, e.g. similarity of word contexts, as well as textual matches in the knowledge base to merge entities together [LME12]. Trani et al. present an approach that is based on manual, collaborative entity linking, and is therefore strongly related to our work [TCL*14]. Moro et al. provide a method that based identified graphs computes entity linking, and go into detail of the possible technical solution to find *similar* entities [MRN14].

All these methods induce considerable extra effort, either when it comes to creating a knowledge-base from Wikipedia, DBPedia, or other sources, as well as computational effort which is problematic for interactive approaches. While our proposed technique could utilize information from knowledge-bases, we do not require any computationally expensive processing or mining steps, as we rely on human knowledge and understanding of the ambiguity resolution and linking task, respectively.

2.3. Entity Relationship Visualization

A number of related work is addressing the problem of visualizing relational data, such as in named entities relationships. In [GFC04], Ghoniem et al. compare the two main approaches, node-link diagrams and (adjacency) matrix visualizations, for their readability. The empirical study found that matrix visualizations are particularly suitable in cases where the associated graph is dense. However, alike the problem of finding a useful 2D layout for nodes in a node-link diagram, matrix-based representations inevitably require an appropriate row-/column ordering for effectively representing the dataset structure. Recently, Behrisch et al. presented in [BBR*16] a survey with guidelines on selecting the most appropriate matrix reordering algorithm. However, the survey states that reordering methods for asymmetric matrices are rare and often lead to unsatisfactory results. In cases, such as the ones presented in this paper, human-assisted reordering should be applied to improve the quality of matrix visualizations [BBR*16, p. 712].

Similar to our proposed visualization, matrix layouts have been applied in various context for showing entity relationships. For example, in MatrixExplorer, Henry and Fekete show a synchronized node-link and matrix exploration interface for depicting entity-relationship structure in social networks [HF06]. A hybrid node-link/matrix visualization, called NodeTrix, was proposed by Henry et al. for representing the co-author relationships in large research communities [HFM07].

Whenever the relationship characteristics are not univariate, but potentially (even conflicting) relationship weights exist, such as in our guiding use case of entity disambiguation and model comparison, node-link diagrams will have the drawback to invest multiple edges for each relationship. On the other hand, matrix-based visualization with a complex glyph cell representations have been introduced by Im et al. in [IML13] and shown, e.g., by Beck and Diehl in [BD10] for the comparative analysis of different dependency relationships in software systems or by Behrisch et al. in [BDS*13] for the comparison of rankings and orderings. Recently, an empirical evaluation of the effectiveness of juxtapositions for triangular matrices has shown that distributions and patterns in large feature spaces can be explored with juxtapositions of asymmetric triangular matrices, showing a different feature/metric on the upper and lower triangular matrix part [LS15].

In the domain of text analysis, entity relationship found specifically appeal. Shen et al. give a dashboard-like view in their tool NameClarifier that allows the simultaneous exploration of different aspects of authors that are subject of disambiguation [SWY*17]. They combine single entity sequences, group-based, and list views to support the disambiguation task. In the tool Jigsaw, Stasko et al. present multiple linked views that deal with entities extracted from documents [SGL08]. Connections between entities are symbolized with lines between multiple lists. To visualize the temporal evolution of entities, Mazeika et al. utilize stacked areas [MTW11]. Oelke et al. present a technique that, based on so-called *fingerprint matrices*, depicts the co-occurrences of entities over the course of

literature [OKK13]. Gold et al. utilize circular visualizations where sectors of the circles represent specific topics or speakers [GREA15]. Their visualization technique supports animation over the temporal progression of the visualized data per entity and shows the corresponding relations popping up over time. Similar, the work of El-Assady et al. utilizes animated display of topics or speakers over time to illustrate their evolution in conversations [EGA*16].

For most of the show related work, we see issues with respect to the scalability of the proposed visualization technique. Common solutions are animations over time, which keeps the number of entities to display simultaneously low. Additionally, techniques that present entities side by side suffer from scalability issues, as the length of the documents, in connection with the chosen aggregation level, dictates the final size of the visualization that could possibly exceed the screen space. Therefore, neither animation nor fingerprint-like visualizations will serve the purpose of display entity information of a whole piece of literature as it is required for ambiguity resolution of them.

3. Interactive Ambiguity Resolution

To support ambiguity resolution of named entities, it is required to give the analyst a way of assessing the relatedness of two given named entities. In our approach, we express relatedness by two measures derived from the analyzed literature: character-level statistics and content-based information. Each of those is represented by a proxy metric, as explained in the next paragraphs.

Character Statistics. We compute the character statistics via so-called pairwise co-occurrences of entities [HAA06]. The reason to compute this particular measure is grounded by the language use in fictional literature. Entities, in particular characters, are commonly referred differently in consecutive sentences, as it is considered to be good writing style. See for example the excerpt from a Harry Potter novel given in Figure 3. There, one entity, Harry Potter, is referenced three times, and each of the references uses a different wording. Co-occurrence metrics are the method to capture those variations in writing style. By utilizing this concept, we are able to express relationships between entities based on their co-occurrence in the form of *Entity A* is co-occurring *n-times* with *Entity B*, and capture variations in language with respect to entities.

Content Information. So far, the occurrences of characters with each other can be quantified and covers, therefore the statistical view on entities. To be able to communicate the content-based context of entities, we compute the word embedding for each entity, that contains its linguistic context [BDVJ03, CW08], and therefore approximates the near content-wise context of an entity. In order to relate the word vectors to entity similarity, the distance of these vectors can be computed by utilizing the Szymkiewicz-Simpson coefficient [Szy34]. Using the resulting similarity score, we are able to compare entities not only based on their co-occurrence, but also on the similarity, or dissimilarity of their context in terms of actual words.

The computation of both metrics is part of Phase Two of the interactive ambiguity resolution process described in the following section.

3.1. Interactive Ambiguity Resolution Process

We structured the interactive ambiguity resolution in a process that consists of four phases. It covers the data generation, model creation, interactive visualization, as well as possible feedback to other processes that makes the interactive ambiguity resolution process suitable to embed in more complex text processing workflows, or act as a stand-alone tool for exploring NER data.

Phase One – Named Entity Recognition. In this phase, named entity recognition software is applied on the documents to process. The result of this phase are the entities and their classes so that the next phases can utilize this information. This phase is executed automatically and does not require any interaction or assistance by users, besides the initial NER technique and the corresponding model selection. We utilize two state of the art and widely used representatives of today's machine learning-based systems, namely the Stanford Named Entity Recognizer [FGM05] and OpenNLP Name Finder (`https://opennlp.apache.org/`) Due to the lack of rule-based NERs, and to be able to cover all NER families given in the Related Work (Section 2), we implemented our own rule-based NER, as described in Section 3.2. It is important to note that at this stage, no assumptions with respect to the technical foundations of named entity recognition as well as the class of entities that are subject to interactive ambiguity resolution are made. Although, as argued before, in this work we concentrate on entities of type person, as we analyze fictional literature that typically relates heavily to characters.

Phase Two – Linguistic Data Preparation. This phase has two purposes: first, it computes the pairwise co-occurrence of entities [HAA06]. We utilize the concept of co-occurrence, as it can be observed that in literature references to the same character in consecutive sentences are expressed differently (see the two examples at the beginning of Section 3.4). Second, the word-embedding [BDVJ03, CW08] per entity is created. Word-embeddings are commonly understood as collections of words that share the same context. We adapted the concept and computed the embedding per entity, which yields a vector of words containing the context of the given entity. Afterward, the pairwise distance per word-vector is computed, which can be interpreted as semantic similarity, because word-embeddings are known to preserve the semantic relationships [MYZ13]. We assume, that a high semantic similarity is a hint for similar entities, and therefore points to potential ambiguities. Similar to Phase Two, this phase is fully automated and does not require interaction. The linguistic data creation is also designed to be as generic as possible, as co-occurrence computation makes no assumptions of the processed text data. Similarly, the creation of the word-embeddings and the resulting semantic similarity measure is possible for all kinds of words, as it poses no requirements to the context it should be computed for.

Phase Three – Interactive Visualization. The third phase contains the depiction of the previous analysis results While the two preceding phases are completely automatized, this phase is where humans and automated data analysis come together, and interactions with the created models from the data analysis are possible. This is demonstrated in Section 4, where we show examples of our

approach executed with three different named entity recognition software packages. We chose the metaphor of an entity-to-entity matrix, where each cell represents a pair of entities. This maps directly to the nature of the data extracted in the two preceding phases, as they are referring to a single entity (Phase One), as well as entity to entity (Phase Two). Details of the visualization and supported interaction are given in Section 3.4 and Section 3.5, respectively.

Phase Four – Feedback. In this work, the primary goal is to improve the NER results, which is prepared in this stage. To do so, interactions such as the merger of two entities are captured (Phase Three) and translated (this phase) into a format that is understood by the NER employed in Phase One. For example, off the shelf modules such as Stanford NER will be accompanied most likely with post-processing based on the user interactions, which is configured and transferred to the corresponding component during this phase. This requires a corresponding post-processing component, as well as a specified way of adding new rules to it. For rule-based NERs, such as our own prototypical implementation, it is possible to formulate new rules and inject them right into the rule set.

For real-world applications, support for the described process can be implemented in various ways. Although, it is clear that the implementation of the visual and interactive processes are not suitable for batch processing of documents, as human interaction is required. Instead, we expect that the outcome of the interactive ambiguity resolution process, which are either postprocessing rules or new additions to rule-based NERs, will be integrated into the data processing pipelines. We also see this process as a tool for interactive model comparison of either different NER methods, or models that differ in parameters important for their training or document selection. The actual configuration depends heavily on the employed NER technology and the use case at hand. Therefore, we see the process elaborated in this section as ideal, and abstracted from use-cases or technical parameters.

3.2. Rule-based Named Entity Detection

In most application areas, rule-based named entity recognition has been superseded by machine-learning based approaches [NS]. Although, there is a big advantage of rule-based approaches: because of the specificity of the typically manually created rule-set, the expected precision is higher than the precision of competing techniques.

A disadvantage of the necessity of a fixed rule-set is the fact, that similar to pre-trained machine-learning models, the input has to conform to the data that has been used to formulate the rules. If this is not the case, rule-sets are expected to fail similarly than model-based techniques, but these errors are easier to counter with extensions of the rule-set. In addition, maintaining a general set of rules is tedious and time-consuming, therefore it makes sense to target a specific type of text, for example, fictional literature, with rule-based NERs. A big advantage of rule-based named entity recognition is the computation time. It is noticeably lower than the processing time of state of the art, model-based NER software. For interactive applications, such as ours, this is a huge plus. Because of these two advantages, expected high precision as well as the suitability for interactive systems, we implemented a custom, rule-based NER software specifically tailored to fictional literature.

Rule	Example
Action Words	**said** Harry **address** Hermione
Possessives	Harry**'s** world Tom**'s** mind
Salutations	**Mr.** Longbottom **Mr.** and **Ms.** Longbottom
Titles	**Professor** Dumbledore **Lord** Voldemort
Verbs, 3rd person, singular	Dudley **walks** towards Harry

Table 2: *The rule-set of our rule-based NER.* ***Bold*** *denotes the main anchors of the rules, red text the extracted entity. The rules are completed with dictionaries that provide* action words*,* salutations*, and* titles*, as well as a set of character-based post-processing rules.*

Table 2 lists the core of the rule-based NER. The rule-set leverages surface properties of the processed text data, which eliminates the need for time and resource intensive parsing of the documents. Additional information, such as dictionaries with titles, salutations, or action words are generated from DBPedia [ABK*07]. Further post-processing is applied on the results so that overlapping matches are included only once, or matches of more than one word are matched as a single entity.

3.3. Automated Analysis Performance

To get an impression of the performance of the three named entity recognition packages (Stanford Named Entity Recognizer, OpenNLP Name Finder, custom rule-based NER), we present the common performance metrics recall and precision and a false positive rate in Table 5. The exact versions and utilized models are given in Table 3. We selected a number of books from our literature collection, so that fictional literature of different length, different authors, and different genres is represented. For practicality, we limited the selection to titles where the creation of a gold standard for the automated evaluation is possible, for example by leveraging various online fan-wikis or Wikipedia.

Software	Version	Notes
Stanford NER	3.6.0	
Stanford NER Models	3.6.0	7 Class, Distsim
OpenNLP Name Finder	1.5.3	
OpenNLP Models	1.5	en-ner-person
Rule-based NER	1.0	
Rule-set	1.0	see Table 2

Table 3: *The utilized named entity recognition software and the corresponding models.*

Table 4 lists the selected books that we show in our examples. To be able to judge the quality of the out of the box performances, we apply common metrics to evaluate named entity recognition software, which are precision and recall. The required gold standard for automated evaluation has been created manually from data sources such as fan-Wikis or Wikipedia. As this is no accepted or widely used data, precision and recall in our studies should not be interpreted absolutely, as the gold standard data contains a number of side characters, animals, or other entities that play only a minor roles. For example, the gold standard for the Harry Potter novel contains *Snowy* and *Tibbles*, which both are cats. These errors were not corrected because we aimed at a wide data foundation of the gold standard. Therefore, the numbers should only serve for relative comparison as they are used in this work, not as absolute performance metrics.

ID	Author and Title	Genre	Year
AC	*Agatha Christie* Death Comes as the End	Mystery	1944
GO	*George Orwell* 1984	Dystopian	1949
JR	*J. K. Rowling* Harry Potter and the Sorcerer's Stone	Fantasy	1997
SK	*Stephen King* Doctor Sleep	Horror	2013

Table 4: *The collection of literature used for our tests.*

Table 5 contains precision, recall and the false positive rate for the chosen literature and the three considered named entity recognition packages. The book from Agatha Christie (AC) seems to be well reflected in the Stanford model, as well as the rule-set of our custom NER. Both technologies reach a recall of 1.00, and compared to the results of analyzing the other books a quite high precision, although the rule-based NER outperforms the other candidates with a precision of 0.87. OpenNLP excels in the analysis of George Orwell's *1984* (GO), the other two NERs have huge problems with this book, in particular, the Stanford NER, that reaches precision and recall values very low compared with the competitors. The results of the analysis of the Harry Potter novel (JR) is similar and almost comparable to all three techniques, while the Stanford NER shows a very good recall compared to OpenNLP and the rule-based NER. For the book of Stephen King (SK), we see that all participants have problems, and perform comparably in terms of the precision, although the recall of OpenNLP and the Stanford NER is better than the recall of the rule-based NER approach. Generally, we see that the false-positive rate of all candidates is quite high, in three of the four displayed cases even over 75 percent, while all techniques have strengths and weaknesses for each of the candidates.

3.4. Entity and Data Visualization

Consider the following sentence: "*Mr. Dursley wondered whether he dared tell her he'd heard the name "Potter"*" [Row97, Chapter One]. While it is a natural assumption, that references to the name *Potter* in a book about *Harry Potter* refer to *Harry Potter*, there is currently no way to actually resolve the reference given in the example sentence, which in the example given above is actually referring to the parents of *Harry Potter*.

Book	NER-Type	Precision	Recall	FP
AC	Stanford	0.41	1.00	0.58
	OpenNLP	0.34	0.57	0.65
	Rules	0.87	1.00	0.12
GO	Stanford	0.01	0.04	0.98
	OpenNLP	0.70	0.80	0.78
	Rules	0.24	0.25	0.76
JR	Stanford	0.23	0.58	0.76
	OpenNLP	0.24	0.27	0.79
	Rules	0.22	0.26	0.77
SK	Stanford	0.09	0.75	0.90
	OpenNLP	0.11	0.51	0.88
	Rules	0.11	0.30	0.88

Table 5: *Precision, recall and the false positive rate of the three NERs. Details about the books can be found in Table 4.*

> "Bless my soul [...] **Harry Potter** ... what an honor." He hurried out from behind the bar, rushed toward **Harry** and seized his hand, tears in his eyes. "Welcome back, **Mr. Potter**, welcome back."

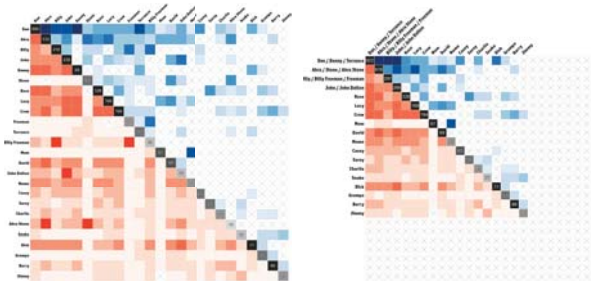
Figure 3: *Harry Potter is referenced in three different ways in three consecutive sentences. Excerpt from Harry Potter and the Sorcerer's Stone by J. K. Rowling [Row97, Chapter Five].*

In the example shown in Figure 3, the same entity, *Harry Potter*, is referenced three times in three sentences, which all use a different wording. With the proposed visual design, we are targeting this exact type of referential ambiguity, which in terms of data analytics can be expressed as the co-occurrence of entities. Admittedly, this technique cannot capture all varieties of possible ambiguities of entity references, as it is based purely on statistics and does not take the contents of the text into account.

To counter this issue, we contrast the co-occurrence information with an additional level of semantic information, the semantic similarity as described in Section 3, description of *Phase Two – Linguistic Data Preparation*. The combination of the semantic similarity measure and co-occurrence information adds the ability to cross-validate findings made either by the inspection of co-occurrences or the semantic similarity so that decisions made on statistics can be backed with semantics and vice versa. This allows findings such as *two entities co-occur together, but do have a different semantic context*, which is a strong pointer that both entities do not refer to the same person, and are not possible by providing only co-occurrence or similarity information.

It is clear, that we need to visualize both types of information in a single view. For each pair of entities, we have two metrics available, namely the co-occurrence and a similarity of the word-embeddings. On data level, this corresponds to an undirected graph (as the scores

are symmetric) of entities with relationships (nodes) that are quantified (weighted) with two numeric values. Visualizing this in a classic node-link diagram is prone for visual clutter, as we have to represent each data value with a separate edge. Instead, we visualize the node-link structure with adjacency matrices, which are known to represent graph structures in space efficient and compact manner. Each cell represents a pair of characters that are indicated by the corresponding rows and columns. In graphs, the neighborships between nodes are symmetric, which in turn leads to symmetric adjacency matrices, which means that in the visualization, two matrix cells are representing the same pair of entities. This property fits well to the data that we want to visualize, as we also have two metrics per entity pair. Recent studies in the area of matrix visualization showed, that asymmetric matrices can be read as effectively as symmetric ones [LS15], so we do not loose interpretability. The following subsections elaborate on the visual design, as well as interaction possibilities of the matrices.



(a) *Initial State* **(b)** *Final State*

Figure 4: *Initial and final state of* AMBIGUITYMATRIX *at the example of Doctor Sleep by Stephen King, generated with the OpenNLP Name Finder.*

3.4.1. Ambiguity Matrix

Frequently co-occurring entities either suggest that individual entities have a high rate of interaction, or that the same entity is addressed in different ways. Visualizing co-occurrences in a matrix helps the user to conceive and to assess entity relationships based on their co-occurrence. The co-occurrences are mapped onto a color scale in which dark and saturated blueish colors represent a high co-occurrence. In the case of no co-occurrence, the cell is white and marked with a gray cross.

To contrast the single viewpoint of entity co-occurrences, we split the matrix into the upper and lower triangular submatrix. The lower-left part of the matrix is used to depict a semantic similarity score based on the entity related word-embedding information, as elaborated in the paragraph *Phase 2 – Linguistic Data Preparation* of Section 3. A different color map is used in the same fashion as for the co-occurrences to emphasize the different attributes, but with reddish colors.

On the diagonal, the overall occurrences of an entity is shown, see Figure 1. This meta-information is represented with a black-to-white color map. For further clarity, the number of occurrences is additionally printed to the cell center.

Having a matrix-based visualization, the ordering or rows and columns plays an important role. In our work, we use an ordering according to the average co-occurrence of each entity. This order typically places the main entities of a book in the upper-left corner of the matrix as they interact with most of the other entity. Entities that are occurring less often, therefore, move more to the lower-right part of the matrix. As the survey in [BBR*16] describes, matrix reordering approaches are mostly designed for symmetric matrix data. However, our visual design implies an asymmetric table ordering. Although correspondent analysis (CA) techniques could be used in this case, we decided to emphasize the semantic and domain-specific row/column ordering, that based on the absolute character occurrences reflects the *importance* of entities in the literature.

3.5. Interacting with Co-occurrences and Semantic Similarity

The user's task is to detect and assess the darker spots in the matrix yielding to a high co-occurrence or a high similarity. The user can now decide to merge the two entities based on the provided information and her background/domain knowledge. Clicking on a cell triggers the merge process for two entities, which is displayed with an animation to help the user to understand the effects in the matrix. Firstly, the row and column belonging to the entity further down and to the right in the matrix are moved towards the entity which is placed more on top and left until they completely overlap. The moving row and column are highlighted during this animation to support the user to track them (see Figure 1b). Afterward, the entity labels at the edges of the matrix are merged, e.g. *Potter / Harry Potter*. Eventually, the merge effects the co-occurrences as well as the similarities to all the other entities. The co-occurrences of the two merge candidates to every other entity are summed up to provide the new co-occurrence value. A recomputation of the similarity score taking into account the merger of two entities is too slow to be performed in appropriate time for the interactive system. Therefore, the maximum similarity score of the merged entities to every other entity adopted, see Figure 1c.

After the entities have been sorted, we do not apply the matrix ordering, as we would certainly damage the navigational context created during the inspection of the matrix by changing the order and therefore also the visual appearance. Instead, we move the merged rows and columns to the top and left, which is a heuristic to keep the sorting mantra intact as all co-occurrences and similarities of the merged character must be equal or higher than the values of the two individual entities. In order to help the user in navigating in the matrix, the current row, column, and their labels are highlighted in the upper and lower part of the matrix whenever the mouse is moved over them, as depicted in Figure 1a. The corresponding cell on the opposite side of the diagonal is also highlighted providing an immediate way to compare co-occurrence and similarity values.

To communicate additional context information with respect to the characters, a tooltip showing additional information about the entities of a hovered cell is displayed, as shown in Figure 5. The tooltip is divided into three parts. On top, the names of the two characters the hovered cell refers to are shown, together with the number of co-occurrences, as well as the similarity score of the character-wise word-embeddings. The second area contains further information about the actual context words, namely the words that are similar and dissimilar from both characters, ranked according

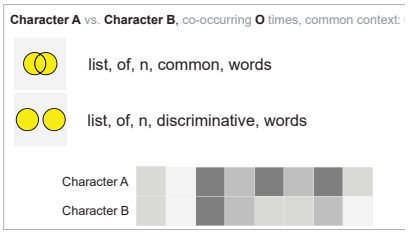


Figure 5: *Illustration of the character cell tooltip. information such as the co-occurrence and the overc contexts is shown. Below, two lists of the ten most imp (top) and distinct (word) are shown. At the bottom, indicate Plutchik's basic emotions [Plu80] of eac intensity, where low indicates light, and dark high ir*

to their word-embedding score. This gives quite goo typical "word-wise" surrounding of a character and insight into the similarities and differences for the cur Last, the area the bottom of Figure 5 illustrates the P emotions [Plu80] per character, extracted with the emotion lexicons by Mohammad and Turney [MT13]. We map the pairwise emotions, which are anger, anticipation, disgust, fear, joy, sadness, surprise and trust, to colored squares. The color is assigned according to the relative intensity, from low (light gray) to high (dark gray). This display communicates the emotional profile of a character, and allows direct comparison between both, helping the user to assess the similarity of their emotional traits.

4. Examples

In this section, we showcase the initial state of AMBIGUITYMATRIX, and the final state after the recognized ambiguities have been resolved. The examples have been conducted by ourselves to keep the results consistent, as we argue that the whole process depends on world-knowledge which differs from user to user, which would make the results almost impossible to discuss.

Agatha Christie: Death Comes as the End As it is visible in Table 5, two of the three NER packages provide already good results with respect to precision, recall, and false positives. It is in line with this insight, that we did not find more than two ambiguous references to entities using the Stanford NER data, as well as the data from the rule-based named entity recognition technique.

George Orwell: 1984 In Figure 2, the initial (left) and final (right) state of the visualization based on data generated with the OpenNLP Name Finder is shown.

In general, and backed by the data given in Table 5, this book seems to be hard to analyze for the Stanford NER, as well as the rule-based NER approach. After loading the data from these two techniques, we saw that no ambiguities were visible for the most frequently occurring 25 characters. From that, we conclude, that most of the errors are caused by rarely occurring characters, which is plausible due to the high false positive rate of all NER techniques. Data created by the OpenNLP Name Finder contained the most resolvable ambiguities in the inspected data (*Winston* and *Winston*

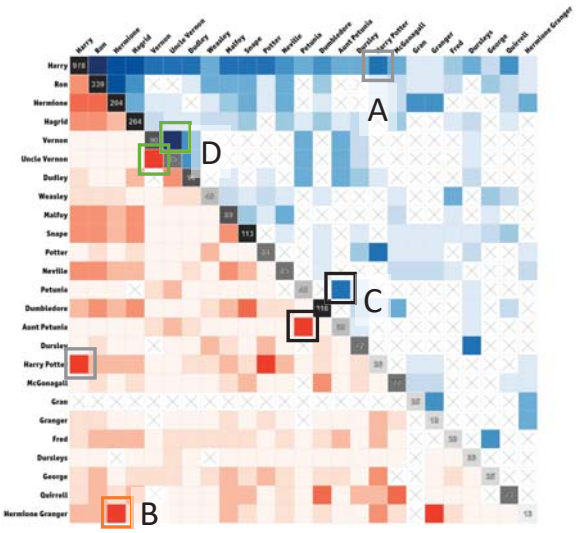


Figure 6: *Initial view of data from the Stanford NER.*

Smith, *Goldstein* and *Emmanuel Goldstein*). This is unexpected because of high precision and recall, although we observe a high false positive rate of 78%, which is likely the cause of these ambiguities.

J. K. Rowling: Harry Potter and the Sorcerer's Stone In Figure 6, the initial view of AMBIGUITYMATRIX with data created by the Stanford NER is shown. Some local patterns stand out in both parts of the matrix, as well as some cells that are immediately recognizable to have either a high co-occurrence or similarity value, indicated by dark blue or red color of the cells. In the figure, we highlighted four of the immediately recognizable ambiguities with colored rectangles as follows: 1. Gray indicates the merger of *Harry* and *Harry Potter*, the co-occurrence as well as the similarity cell indicate high overlap of these two entities (A); 2. The ambiguity of *Hermione* and *Hermione Granger* stands out in the similarity part of the matrix, where it is indicated with an orange rectangle (B); 3. The consolidation of *Petunia* and *Aunt Petunia* is indicated with black (C); 4. Another case where the salutation is part of the consolidation and therefore resolves the ambiguity between *Vernon* and *Uncle Vernon* in the analysis results is highlighted with green rectangles (D).

Because of the similar performance of the other named entity recognition tools, their findings are also similar. Some details differ, for example, the OpenNLP-based results do not suggest to merge *Harry* and *Potter*. The rule-based system seems to be stronger with entities containing salutations. In our experiment, the visualization based on rule-based NER data was the only one that prompted to consider *Dumbledore* and *Professor Dumbledore* as a single entity.

Stephen King: Doctor Sleep In the last example, we applied AMBIGUITYMATRIX on the horror novel *Doctor Sleep* by Stephen King.

The initial and final state, created with data generated by the rule-based NER, is depicted in Figure 4. We found a number of clear ambiguities, such as or *Abra* and *Abra Stone*, that are shown

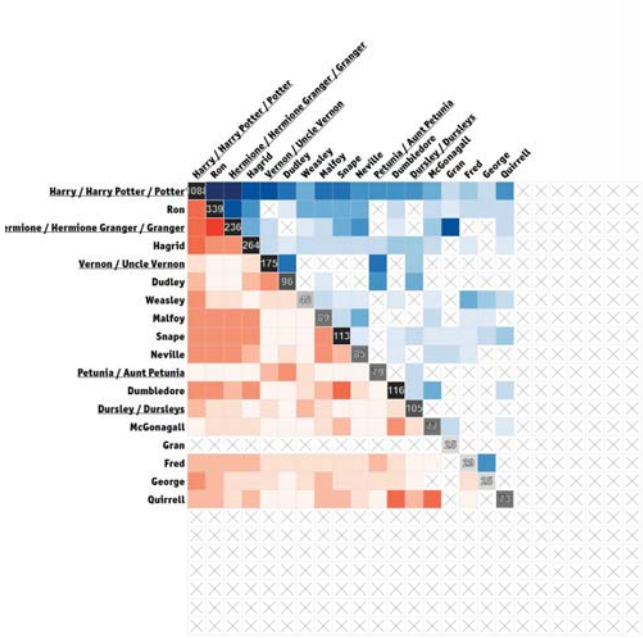


Figure 7: *Final view of data from the Stanford NER that started with 25 entities. This view shows 18 entities.*

in Figure 4b. Noteworthy is, that in contrast to the other candidates, we found some examples where the similarity, as well as the co-occurrence do not correspond towards merging characters, such as it is the case for *Dan* and *Danny*. While the co-occurrence is pointing in the direction of a candidate of ambiguities to merge, the similarity does not stand out. This is one of the cases, where human knowledge is required to make sure that the two candidates are merged for good reason.

An interesting observation made during the ambiguity resolution process is, that with the merger and immediate display of the results, transitive ambiguities, and their possible consolidation get visible. In the example of the Harry Potter novel, after merging *Harry* and *Harry Potter*, the connection of the newly merged entity and *Potter* got clearly visible. This lead to an entity composed out of *Harry*, *Harry Potter*, and *Potter*. A similar observation can be made with *Hermione*, *Hermione Granger*, and *Granger*.

5. Authorship Profiling using AMBIGUITYMATRIX

While our work is focused on guiding human users in the non-trivial process of resolving ambiguity problems of state of the art natural language processing methods used to detect and classify entities of class *person*, we also found that AMBIGUITYMATRIX can be used for authorship profiling for the examples of our literature collection. During our experiments with the available literature that comprises more than 20 different authors, we found that the co-occurrence view, as provided by AMBIGUITYMATRIX, allows distinguishing different authors. In Figure 8 we showcase examples from our experiments, where differences of the authors are clearly visible.

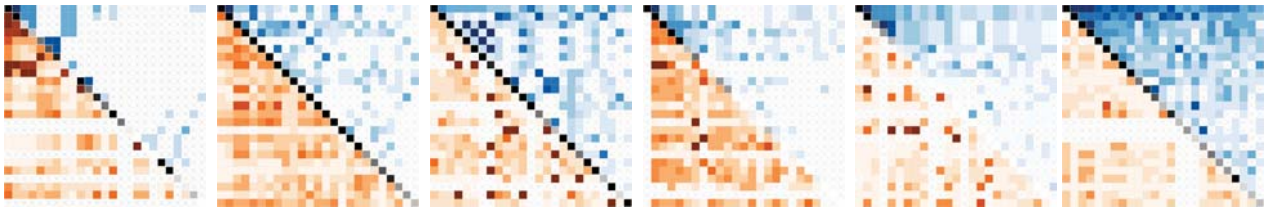
In Figure 8a, *Tales of the Grotesque and Arabesque* by Edgar Allan Poe shows unique patterns, both in the co-occurrence and semantic similarity portion of the matrix. There is a low number of interacting entities, and a degree of semantic similarity is shared by the top entities can be observed that is not showing up in any of the other presented examples in Figure 8. The visualization of Tolkiens *The Two Towers* reveals only a very small number of co-occurrences and corresponding entities, high semantic similarity is limited to a small number of entities, too (Figure 8b). For Agatha Christie, a number of small, grid-like co-occurrence pattern are a unique characteristic (Figure 8c). In addition, compared to the other books, there are many entity pairs with high semantic similarity. Stephen Kings *The Shining* shows overall a low degree of entity co-occurrences, also a group pattern can be seen on the top left of the visualization in Figure 8d. In Figure 8e, the visualization of *Angels and Daemons* by Dan Brown shows an outstanding pattern, too. The upper part of the co-occurrence view shows groups of entities that co-occur, which cannot be observed in the visualizations of the other books. Also, compared to the other examples, the degree of semantic similarity between the different entities is quite low. Last, the visualization in Figure 8f is showing a degree of entity co-occurrence that is outstanding in the presented examples. In contrast, the semantic similarity seems to be low overall the entities.

These observations give a strong hint that examining the co-occurrence of entities, in combination with a semantic similarity of entities, for the purpose of author profiling could be valuable, as current technologies are typically based on letter-level, lexical, syntactic or semantic information and do not take entity-based, content-related text features into account [Sta09].

6. Discussion and Future Perspectives

In addition to what has been shown in this work, iterations with adapted NER models or rule-sets are also possible, for example by displaying more than one visualization in juxtaposed views. This comparison has also potential to allow the visual examination of the model evolution. Matrix-based visualizations are specifically powerful for depicting visual patterns. In a small-multiple setting, the presence and absence of visual patterns caused by the corresponding interactions of the user can be spotted easily. We are currently experimenting with these user-interface extensions, which are additionally providing a graphical depiction of analytic provenance, similar to the illustration in Figure 1.

Coming back to the presented use-case of interactive named entity disambiguation, we introduced our work as a tool to create (post-processing) rules. It is natural to extend our approach feeding the user interactions back to the named entity recognition process, in order to adapt the corresponding data analysis techniques. For a rule-based NER, an inference of new rules based on the manually selected entities is possible. For machine learning-based techniques, a large number of interactions and examples will be needed in order to be able to re-train the model or the employed method to incorporate the human feedback. The available models are trained on huge data sets, and a small number of new examples will have no noticeable impact on the performance. To cope with that problem, techniques to over-represent the manually selected examples could be employed. For interactive systems, this can lead to significant processing and



(a) *Edgar Allan Poe: Tales of the Grotesque and Arabesque* **(b)** *J. R. R. Tolkien: The Two Towers* **(c)** *Agatha Christie: A Caribbean Mystery* **(d)** *Stephen King: The Shining* **(e)** *Dan Brown: Angels and Daemons* **(f)** *J. K. Rowling: Harry Potter and the Order of the Phoenix*

Figure 8: *Illustration of visual author profiling by* AMBIGUITYMATRIX.

waiting times for machine-learning based NER software, though this problem will not occur for rule-based NER.

While we focus on *semantic ambiguities* that are grounded in typical errors such as missing salutations or family names, in some cases we found errors and duplicates caused by punctuation and other characters, that have not been stripped from the extracted entity. This allows the distinction of four types of meaningful rules that are possible to generate with AMBIGUITYMATRIX. First, rules based on ambiguities of entities indicated by co-occurrence in the upper right part of AMBIGUITYMATRIX. Second, the semantic similarity, as depicted in the lower left part of the matrix visualization. Third, there are rules which are triggered by both, the co-occurrence and semantic similarity. And fourth, rules that correct spelling based post-processing. This information could be a good start for future extensions of the presented analysis and visualization technique.

The presented approach was designed for the interactive improvement of named entity recognition but can also serve other purposes, such as interactive model comparisons, which can be done by showing multiple of the proposed visualization side by side. Enriching the views with linking and brushing is an effective tool for comparisons, in particular of the data represented per column and row is kept the same for all views. Besides different models and software packages, also different parameter sets can be compared interactively.

In this work, we concentrated on the entities of type *person*, while NER software typically assigns more classes, such as *location*, *company* or *time*. It is of interest to examine whether the proposed methodology is also applicable for these entity classes, although, we suspect this is not the case. The character disambiguation of characters (persons) as presented in this work relies heavily on the type of referral to the characters, because of the language used by the author, see Figure 3 for an example. From our experience with the data, this cannot be assumed for other entity classes, such as locations or company names. Location or companies are commonly addressed by their name. In consequence, we do not expect to find enough variance in the data, so that the idea of exploiting co-occurrences to point to possible candidates for a merger will work. The semantic similarity based on word-embeddings should still be useful, as the embedding does not depend on the word it is created for.

During our experiments, we found also misleading cues. For example, in the Harry Potter novel that is subject to the Example Section 4, the visual hint was pointing strongly to merge *Professor McGonagall* and *Professor Dumbledore*. The semantic similarity between these two characters is 0.83, in a range from 0 to 1, while the visualization of co-occurrence was not strongly indicating a possible merging candidate. While these problematic cases were only rarely observed by us, they are a hint to add even more context as presented in Section 3.5 to the visualization, for example by presenting relevant text snippets of entity (co-) occurrences.

In this work, all matrices presented contain at max 25 characters. We did this on purpose, as experiments with the data showed that books in our collection did not contain more than 25 main characters, determined by frequent interaction. For this work, there are two different types of scalability to take into account: data-, and visual scalability. Data-wise, increasing the number of characters is not an issue, as the employed NER toolkits already generate more than 25 characters. In a collection of around 1,600 e-books, we found that the majority of books contains between one and 200 characters (entities of type person), while the number of character co-occurring with others is much lower, around 120. A study by Ghoniem et al. [GFC04] suggests that even for 100 characters, the matrix visualization stays interpretable, and will perform better than competing techniques such as node-link diagrams. Therefore, we argue that the proposed visualization technique meets the scalability requirements for the presented task.

7. Conclusion

In this paper, we introduced an interactive, visualization-based technique to resolve ambiguities in named entity recognition, tailored to entities that are classified as *persons*, which in the context of our work maps to characters in fictional literature. The corresponding process has been separated into four phases, which all show generalization potential and adaptability for further use-cases beyond the one presented in this work. We understand our work as a step in the direction of making the complex problem of named entity recognition easier to improve by involving a human analyst, as today's automated methods, even though they have an extensive support of machine-readable sources of knowledge, cannot grasp an author's intention as humans arguably can.

Acknowledgments

This work was supported by the EU project Visual Analytics for Sense-making in Criminal Intelligence Analysis (VALCRI) under grant number FP7-SEC-2013-608142.

References

[ABK*07] AUER S., BIZER C., KOBILAROV G., LEHMANN J., CYGANIAK R., IVES Z. G.: Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007.* (2007), pp. 722–735. URL: http://dx.doi.org/10.1007/978-3-540-76298-0_52, doi:10.1007/978-3-540-76298-0_52. 5

[BBR*16] BEHRISCH M., BACH B., RICHE N. H., SCHRECK T., FEKETE J.: Matrix reordering methods for table and network visualization. *Comput. Graph. Forum 35*, 3 (2016), 693–716. URL: http://dx.doi.org/10.1111/cgf.12935, doi:10.1111/cgf.12935. 3, 7

[BD10] BECK F., DIEHL S.: Visual comparison of software architectures. *Proceedings of the 5th international symposium on Software visualization - SOFTVIS '10* (2010), 183. doi:10.1145/1879211.1879238. 3

[BDS*13] BEHRISCH M., DAVEY J., SIMON S., SCHRECK T., KEIM D. A., KOHLHAMMER J.: Visual Comparison of Orderings and Rankings. In *EuroVis Workshop on Visual Analytics* (2013), Pohl M., Schumann H., (Eds.), The Eurographics Association, pp. 1–7. doi:10.2312/PE.EuroVAST.EuroVA13.007-011. 3

[BDVJ03] BENGIO Y., DUCHARME R., VINCENT P., JANVIN C.: A neural probabilistic language model. *Journal of Machine Learning Research 3* (2003), 1137–1155. URL: http://www.jmlr.org/papers/v3/bengio03a.html. 2, 4

[BMS00] BALUJA S., MITTAL V. O., SUKTHANKAR R.: Applying machine learning for high-performance named-entity extraction. *Computational Intelligence 16*, 4 (2000), 586–596. URL: http://dx.doi.org/10.1111/0824-7935.00129, doi:10.1111/0824-7935.00129. 2

[BMSW97] BIKEL D. M., MILLER S., SCHWARTZ R. M., WEISCHEDEL R. M.: Nymble: a high-performance learning name-finder. In *ANLP* (1997), pp. 194–201. URL: http://aclweb.org/anthology-new/A/A97/A97-1029.pdf. 2

[BON03] BENDER O., OCH F. J., NEY H.: Maximum entropy models for named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003* (2003), pp. 148–151. URL: http://aclweb.org/anthology/W/W03/W03-0420.pdf. 2

[BP06] BUNESCU R. C., PASCA M.: Using encyclopedic knowledge for named entity disambiguation. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy* (2006). URL: http://acl.ldc.upenn.edu/E/E06/E06-1002.pdf. 2, 3

[BSAG98] BORTHWICK A., STERLING J., AGICHTEIN E., GRISHMAN R.: Nyu: Description of the mene named entity system as used in muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998* (1998). 2

[CM99] CALIFF M. E., MOONEY R. J.: Relational learning of pattern-match rules for information extraction. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence, July 18-22, 1999, Orlando, Florida, USA.* (1999), pp. 328–334. URL: http://www.aaai.org/Library/AAAI/1999/aaai99-048.php. 2

[CW08] COLLOBERT R., WESTON J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008* (2008), pp. 160–167. URL: http://doi.acm.org/10.1145/1390156.1390177, doi:10.1145/1390156.1390177. 2, 4

[DC08] DIESNER J., CARLEY K. M.: Conditional random fields for entity extraction and ontological text coding. *Computational & Mathematical Organization Theory 14*, 3 (2008), 248–262. URL: http://dx.doi.org/10.1007/s10588-008-9029-z, doi:10.1007/s10588-008-9029-z. 2

[DNOO08] DEN Y., NAKAMURA J., OGISO T., OGURA H.: A proper approach to japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco* (2008). URL: http://www.lrec-conf.org/proceedings/lrec2008/summaries/258.html. 2

[dSKL04] DA SILVA J. F., KOZAREVA Z., LOPES J. G. P.: Cluster analysis and classification of named entities. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal* (2004). URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/796.pdf. 2

[EGA*16] EL-ASSADY M., GOLD V., ACEVEDO C., COLLINS C., KEIM D. A.: Contovi: Multi-party conversation exploration using topic-space views. *Comput. Graph. Forum 35*, 3 (2016), 431–440. URL: http://dx.doi.org/10.1111/cgf.12919, doi:10.1111/cgf.12919. 4

[FG15] FLEKOVA L., GUREVYCH I.: Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015* (2015), pp. 1805–1816. URL: http://aclweb.org/anthology/D/D15/D15-1208.pdf. 2

[FGM05] FINKEL J. R., GRENAGER T., MANNING C. D.: Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA* (2005). URL: http://acl.ldc.upenn.edu/P/P05/P05-1045.pdf. 4

[GFC04] GHONIEM M., FEKETE J., CASTAGLIOLA P.: A comparison of the readability of graphs using node-link and matrix-based representations. In *10th IEEE Symposium on Information Visualization (InfoVis 2004), 10-12 October 2004, Austin, TX, USA* (2004), pp. 17–24. URL: http://dx.doi.org/10.1109/INFVIS.2004.1, doi:10.1109/INFVIS.2004.1. 3, 10

[GREA15] GOLD V., ROHRDANTZ C., EL-ASSADY M.: Exploratory Text Analysis using Lexical Episode Plots. In *Eurographics Conference on Visualization (EuroVis) - Short Papers* (2015), Bertini E., Kennedy J., Puppo E., (Eds.), The Eurographics Association. doi:10.2312/eurovisshort.20151130. 4

[HAA06] HASSELL J., ALEMAN-MEZA B., ARPINAR I. B.: Ontology-driven automatic entity disambiguation in unstructured text. In *The Semantic Web - ISWC 2006, 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, November 5-9, 2006, Proceedings* (2006), pp. 44–57. URL: http://dx.doi.org/10.1007/11926078_4, doi:10.1007/11926078_4. 4

[HF06] HENRY N., FEKETE J.: Matrixexplorer: a dual-representation system to explore social networks. *IEEE Trans. Vis. Comput. Graph. 12*, 5 (2006), 677–684. URL: http://dx.doi.org/10.1109/TVCG.2006.160, doi:10.1109/TVCG.2006.160. 3

[HFM07] HENRY N., FEKETE J., MCGUFFIN M. J.: Nodetrix: a hybrid visualization of social networks. *IEEE Trans. Vis. Comput. Graph. 13*, 6 (2007), 1302–1309. URL: http://dx.doi.org/10.1109/TVCG.2007.70582, doi:10.1109/TVCG.2007.70582. 3

[HYB*11] HOFFART J., YOSEF M. A., BORDINO I., FÜRSTENAU H., PINKAL M., SPANIOL M., TANEVA B., THATER S., WEIKUM G.: Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL* (2011), pp. 782–792. URL: http://www.aclweb.org/anthology/D11-1072. 3

[IML13] IM J., MCGUFFIN M. J., LEUNG R.: GPLOM: the generalized plot matrix for visualizing multidimensional multivariate

data. *IEEE Trans. Vis. Comput. Graph. 19*, 12 (2013), 2606–2614. URL: http://dx.doi.org/10.1109/TVCG.2013.160, doi:10.1109/TVCG.2013.160. 3

[LME12] LIN T., MAUSAM, ETZIONI O.: Entity linking at web scale. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction* (Stroudsburg, PA, USA, 2012), AKBC-WEKEX '12, Association for Computational Linguistics, pp. 84–88. URL: http://dl.acm.org/citation.cfm?id=2391200.2391216. 3

[LS15] LIU X., SHEN H.: The effects of representation and juxtaposition on graphical perception of matrix visualization. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015* (2015), pp. 269–278. URL: http://doi.acm.org/10.1145/2702123.2702217, doi:10.1145/2702123.2702217. 3, 7

[LZ12] LIU B., ZHANG L.: A survey of opinion mining and sentiment analysis. In *Mining Text Data*, Aggarwal C. C., Zhai C., (Eds.). Springer US, Boston, MA, 2012, pp. 415–463. URL: http://dx.doi.org/10.1007/978-1-4614-3223-4_13, doi:10.1007/978-1-4614-3223-4_13. 2

[MBK14] MAKAZHANOV A., BARBOSA D., KONDRAK G.: Extracting family relationship networks from novels. *CoRR abs/1405.0603* (2014). URL: http://arxiv.org/abs/1405.0603. 2

[MC07] MIHALCEA R., CSOMAI A.: Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007* (2007), pp. 233–242. URL: http://doi.acm.org/10.1145/1321440.1321475, doi:10.1145/1321440.1321475. 3

[MRN14] MORO A., RAGANATO A., NAVIGLI R.: Entity linking meets word sense disambiguation: a unified approach. *TACL 2* (2014), 231–244. URL: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/291. 3

[MT13] MOHAMMAD S. M., TURNEY P. D.: Crowdsourcing a wordâĂŞemotion association lexicon. *Computational Intelligence 29*, 3 (2013), 436–465. URL: http://dx.doi.org/10.1111/j.1467-8640.2012.00460.x, doi:10.1111/j.1467-8640.2012.00460.x. 8

[MTW11] MAZEIKA A., TYLENDA T., WEIKUM G.: Entity timelines: visual analytics and named entity evolution. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011* (2011), pp. 2585–2588. URL: http://doi.acm.org/10.1145/2063576.2064026, doi:10.1145/2063576.2064026. 3

[MYZ13] MIKOLOV T., YIH W., ZWEIG G.: Linguistic regularities in continuous space word representations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA* (2013), pp. 746–751. URL: http://aclweb.org/anthology/N/N13/N13-1090.pdf. 4

[NC14] NEELAKANTAN A., COLLINS M.: Learning dictionaries for named entity recognition using minimal supervision. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden* (2014), pp. 452–461. URL: http://aclweb.org/anthology/E/E14/E14-1048.pdf. 3

[NS] NADEAU D., SEKINE S.: A survey of named entity recognition and classification. 3–26. doi:10.1075/bct.19.03nad. 1, 3, 5

[OKK13] OELKE D., KOKKINAKIS D., KEIM D. A.: Fingerprint matrices: Uncovering the dynamics of social networks in prose literature. *Computer Graphics Forum 32*, 3 (2013), 371–380. URL: http://dx.doi.org/10.1111/cgf.12124, doi:10.1111/cgf.12124. 4

[Plu80] PLUTCHIK R.: A general psychoevolutionary theory of emotion. *Theories of emotion 1*, 3-31 (1980), 4. 8

[Rau91] RAU L. F.: Extracting company names from text. In *[1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application* (Feb 1991), vol. i, pp. 29–32. doi:10.1109/CAIA.1991.120841. 2

[RC12] READ J., CARROLL J. A.: Annotating expressions of appraisal in english. *Language Resources and Evaluation 46*, 3 (2012), 421–447. URL: http://dx.doi.org/10.1007/s10579-010-9135-7, doi:10.1007/s10579-010-9135-7. 2

[Row97] ROWLING J. K.: *Harry Potter and the sorcerer's stone*. Scholastic, 1997. 6

[Sar08] SARAWAGI S.: Information extraction. *Found. Trends databases 1*, 3 (Mar. 2008), 261–377. URL: http://dx.doi.org/10.1561/1900000003, doi:10.1561/1900000003. 3

[SGL08] STASKO J. T., GÖRG C., LIU Z.: Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization 7*, 2 (2008), 118–132. URL: http://dx.doi.org/10.1057/palgrave.ivs.9500180, doi:10.1057/palgrave.ivs.9500180. 3

[SKKS01] SEON C., KO Y., KIM J., SEO J.: Named entity recognition using machine learning methods and pattern-selection rules. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, November 27-30, 2001, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan* (2001), pp. 229–236. URL: http://www.afnlp.org/nlprs2001/pdf/0143-01.pdf. 3

[SR09] SEKINE S., RANCHHOD E. (Eds.): *Named Entities: Recognition, classification and use*. John Benjamins Publishing Company, jul 2009. URL: http://dx.doi.org/10.1075/bct.19, doi:10.1075/bct.19. 2

[SS04] SHINYAMA Y., SEKINE S.: Named entity discovery using comparable news articles. In *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland* (2004). URL: http://www.aclweb.org/anthology/C04-1122. 2

[Sta09] STAMATATOS E.: A survey of modern authorship attribution methods. *JASIST 60*, 3 (2009), 538–556. URL: http://dx.doi.org/10.1002/asi.21001, doi:10.1002/asi.21001. 9

[SWY*17] SHEN Q., WU T., YANG H., WU Y., QU H., CUI W.: Nameclarifier: A visual analytics system for author name disambiguation. *IEEE Trans. Vis. Comput. Graph. 23*, 1 (2017), 141–150. URL: http://doi.ieeecomputersociety.org/10.1109/TVCG.2016.2598465, doi:10.1109/TVCG.2016.2598465. 3

[Szy34] SZYMKIEWICZ D.: *Une contribution statistique a la géographie floristique*. Polskie Towarzystwo Botaniczne, 1934. 4

[TCL*14] TRANI S., CECCARELLI D., LUCCHESE C., ORLANDO S., PEREGO R.: Manual annotation of semi-structured documents for entity-linking. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014* (2014), pp. 2075–2077. URL: http://doi.acm.org/10.1145/2661829.2661854, doi:10.1145/2661829.2661854. 3

[VJPR15] VALA H., JURGENS D., PIPER A., RUTHS D.: 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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015* (2015), pp. 769–774. URL: http://aclweb.org/anthology/D/D15/D15-1088.pdf. 2