Fact Distribution in Information Extraction

Mark Stevenson
Department of Computer Science
Regent Court, 211 Portobello Street,
University of Sheffield
Sheffield
S1 4DP, UK
marks@dcs.shef.ac.uk

January 2, 2007

Abstract. Several recent Information Extraction (IE) systems have been restricted to the identification facts which are described within a single sentence. It is not clear what effect this has on the difficulty of the extraction task or how the performance of systems which consider only single sentences should be compared with those which consider multiple sentences. This paper compares three IE evaluation corpora, from the Message Understanding Conferences, and finds that a significant proportion of the facts mentioned therein are not described within a single sentence. Therefore systems which are evaluated only on facts described within single sentences are being tested against a limited portion of the relevant information in the text and it is difficult to compare their performance with other systems. Further analysis demonstrates that anaphora resolution and world knowledge are required to combine information described across multiple sentences. This result has implications for the development and evaluation of IE systems.

Keywords: Information Extraction, evaluation, Message Understanding Conferences

1. Introduction

Information Extraction (IE) is the process of identifying specific pieces of information in text, for example, the movements of company executives or the victims of terrorist attacks. IE is a complex task; information may be spread across a document. Several sentences or paragraphs of a text may have to be examined to identify a fact. For example, the following two sentences describe a management succession event (i.e. a change in corporate management personnel):

> Pace American Group Inc. said it notified two top executives it intends to dismiss them because an internal investigation found evidence of "self-dealing" and "undisclosed financial relationships." The executives are Don H. Pace, cofounder, president and chief executive officer; and Greg S. Kaplan, senior vice president and chief financial officer.

The fact that the executives are leaving and the name of the organisation are listed in the first sentence while the names of the executives

and their posts are listed in the second sentence. The succession events can only be fully understood from a combination of the information contained in both sentences.

Combining the required information from multiple sentences is not a simple task since it is necessary to identify phrases which refer to the same entities, "two top executives" and "the executives" in the above example. Additional difficulties occur because the same entity may be referred to in different ways. For example, "International Business Machines Ltd." may be referred to by an abbreviation ("IBM"), nickname ("Big Blue") or an anaphoric expression such as "it" or "the company". These complications make it difficult to identify the correspondences between different portions of the text describing this event.

Traditionally IE systems have consisted of several components with some analysing each sentence and others being responsible for combining the information discovered (Grishman, 2003). These systems were often designed for a specific extraction task and could only be modified by experts. In an effort to overcome this brittleness machine learning methods have been applied to port systems to new domains and extraction tasks with minimal manual intervention. However, many of these systems consider each sentence in isolation and only extract facts which are described within a single sentence, examples include (Soderland, 1999; Yangarber et al., 2000; Chieu and Ng, 2002; Zelenko et al., 2003; Culotta and Sorensen, 2004; Stevenson and Greenwood, 2005; Sekine, 2006). For the remainder of this paper we shall refer to these systems as Single Sentence Approaches (SSA). Conversely, IE systems that have the ability to identify information described in more than one sentence are Multiple Sentence Approaches (MSA).

The development of SSA systems is now a well established methodology in IE research. However, since SSA systems analyse each sentence in isolation and do not attempt to combine items from different sentences, they are limited to identifying information described within a single sentence but are unable to recognise facts expressed across multiple sentences. In the above example these systems could recognise the fact that one of Pace's job titles is "president" but not that he is employed by Pace America Group. Of course, this relation could be identified using techniques for combining information across sentences but this is rarely applied; none of the cited examples of SSA systems use anaphora resolution to help identify relations between items mentioned in different sentences.

A possible reason for developing SSA systems may be the assumption that the majority of facts described in the text are expressed within a single sentence and there is little to be gained from the extra processing required to combine information. In fact, SSA systems only

report results across the facts they consider i.e. those expressed within a single sentence. Conversely MSA systems consider a wider set of facts and report results across those. The facts considered by SSA systems are a subset of those examined by MSA systems but the proportion is not known, making it difficult to compare their performance.

This paper describes an analysis of three corpora commonly used to evaluate IE systems which demonstrates that a significant proportion (up to 60%) of the facts in those documents cannot be identified by SSA systems. The fact that documents contains facts described across multiple sentences is not surprising in itself but it might not be expected that such a large proportion fall into this category. This result demonstrates that SSA systems do not properly consider a large proportion of facts in text and this has implications for the evaluation and development of IE systems. Results of SSA systems which have been previously reported cannot be considered to be directly comparable with those for MSA approaches and should be reinterpreted. In addition, IE system designers who use SSA techniques cannot expect their systems to identify all facts within texts and effort must be spent on the development of techniques for extracting facts described across multiple sentences.

The remainder of this paper is organised as follows. Section 2 describes data from the Message Understanding Conferences and Section 3 the process that is applied to it to determine the proportion of facts they contain which are described in a single sentence. Section 4 describes the results of this analysis. Section 5 discusses the ways in which facts described across multiple sentences could be identified and describes an experiment which estimates the amount of additional facts which could be found if anaphora resolution was applied. Section 6 summarises related work while Section 7 discusses the implications which can be drawn from this analysis.

2. MUC Templates

The data used for the experiments described in this paper are taken from various Message Understanding Conferences (MUCs). These were a series of seven conferences run between 1987 and 1998 which were intended to evaluate the accuracy of IE systems. The evaluation regime gradually evolved over the course of the conference series but always followed the same general format. An IE task was defined and participants provided with sample documents describing information pertinent to the task along with completed templates demonstrating what should be extracted from them. Participants are given a period of time to develop their systems to carry out the extraction task. At the end of this time

each system is evaluated by running it over the evaluation documents and its results compared against manually completed templates which the participants do not have access to. A conference is then held to discuss the results and their implications.

The experiments described here make use of the evaluation data from three of the MUC conferences: (1) MUC4, for which the extraction task was concerned with reports of terrorist incidents in Latin America, (2) MUC6, dealt with management succession events and (3) MUC7, rocket launches. These corpora are commonly used to evaluate IE systems. The aim of these evaluation was to develop systems which could fill answer templates with information extracted from text. The templates consisted of three basic elements: **String Slots** which are filled using strings extracted directly from the text; **Text Conversion Slots** and **Set Fill Slots** which contain values that have to be inferred for the document. Figure 1 shows a filled template from the MUC4 evaluation. Slots 9 and 10 are examples of string slots. These often list alternative expressions which refer to the same entity, such as "FARABUNDO MARTI NATIONAL LIBERATION FRONT" and "FMLN" for slot 10. Slots 4 and 5 are text conversion slots which are completed using a set of pre-defined values for each slot. Slots 14 and 21 are set fill slots. These are completed by deriving the number of items falling into a particular class and enumerating that list. After MUC4 a more complex nested template structure was adopted for subsequent evaluations. This new structure effectively retained the use of three basic slot types.

3. Fact Matching

Our goal is to identify the proportion of facts in the MUC corpora which are described within a sentence and can therefore be extracted by a SSA IE system. One way to estimate this is to examine the template's string slots, which are taken directly from the text, and compute the proportion of events for which all of the string slots occur within the same sentence. The rationale behind this approach is that, since each sentence is examined separately, items much occur together for these approaches to identify the connection between them. This procedure will provide an upper bound on the number of facts which are described within one sentence; if the string slots cannot be found together in a sentence then that event must be described across multiple sentences but, on the other hand, if they do occur in the same sentence then that does not necessarily mean the event is described therein. This approach only considers string slots in the template and this is because

0. MESSAGE: ID	DEV-MUC3-0190 (ADS)
1. MESSAGE: TEMPLATE	2
2. INCIDENT: DATE	- 26 APR 89
3. INCIDENT: LOCATION	EL SALVADOR: SAN SALVADOR (CITY): SAN MIGUELITO (NEIGHBORHOOD)
4. INCIDENT: TYPE	BOMBING
5. INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
6. INCIDENT: INSTRUMENT ID	‘‘BOMB’’
7. INCIDENT: INSTRUMENT TYPE	BOMB: ‘‘BOMB’’
8. PERP: INCIDENT CATEGORY	TERRORIST ACT
9. PERP: INDIVIDUAL ID	‘‘URBAN GUERRILLA GROUP’’
10. PERP: ORGANIZATION ID	‘‘FARABUNDO MARTI NATIONAL LIBERATION FRONT’’ / ‘‘FMLN’’
11. PERP: ORGANIZATION CONFIDENCE	POSSIBLE: ‘‘FARABUNDO MARTI NATIONAL LIBERATION FRONT’’ / ‘‘FMLN’’
12. PHYS TGT: ID	‘‘ARMORED VEHICLE’’
13. PHYS TGT: TYPE	TRANSPORT VEHICLE: ‘‘ARMORED VEHICLE’’
14. PHYS TGT: NUMBER	1: ‘‘ARMORED VEHICLE’’
15. PHYS TGT: FOREIGN NATION	-
16. PHYS TGT: EFFECT OF INCIDENT	-
17. PHYS TGT: TOTAL NUMBER	-
18. HUM TGT: NAME	‘‘ROBERTO GARCIA ALVARADO’’
19. HUM TGT: DESCRIPTION	‘‘ATTORNEY GENERAL’’: ‘‘ROBERTO GARCIA ALVARADO’’
20. HUM TGT: TYPE	GOVERNMENT OFFICIAL / LEGAL OR JUDICIAL: ‘‘ROBERTO GARCIA ALVARADO’’
21. HUM TGT: NUMBER	1: ‘‘ROBERTO GARCIA ALVARADO"
22. HUM TGT: FOREIGN NATION	-
23. HUM TGT: EFFECT OF INCIDENT	DEATH: ‘‘ROBERTO GARCIA ALVARADO’’
24. HUM TGT: TOTAL NUMBER	-

AS THE PRESIDENT-ELECT WAS MAKING THIS STATEMENT, HE LEARNED ABOUT THE ASSASSINATION OF ATTORNEY GENERAL ROBERTO GARCIA ALVARADO. [SENTENCE AS PUBLISHED] ALVARADO WAS KILLED BY A BOMB PRESUMABLY PLACED BY AN URBAN GUERRILLA GROUP ON TOP OF HIS ARMORED VEHICLE AS IT STOPPED AT AN INTERSECTION IN SAN MIGUELITO NEIGHBORHOOD, NORTH OF THE CAPITAL.

Figure 1. Example MUC4 template and text from which it was extracted

it is straightforward to identify where they are mentioned in text but difficult to complete automatically for other fields.

The matching process was applied to the evaluation corpora used for three of the MUC exercises as follows: the text was initially split into sentences.[1] The set of possible fillers were then extracted from

[1] The MUC6 and MUC7 texts were split into sentences using the Edinburgh University LT-TTT tool (Grover et al., 2000) . The MUC4 texts are written entirely in upper case and were split using a version of the OpenNLP tools sentence detec-

the answer key templates and converted into a regular expression for pattern matching.[2] Each fact was then compared against each sentence in the document it was derived from and the sentence for which the most fields matched stored.

As part of this matching process the facts in each MUC template are transformed into a common representation which includes the most important information contained in the string slots and makes the process of comparing each fact against the text more straightforward. The common representation consists of a set of fields each of which has at least one associated filler. For example, the following event consists of three fields which have, respectively, three, two and one possible fillers, separated by "|":

```
PERP(SHINING PATH MEMBERS|SHINING PATH|
    150 SHINING PATH MEMBERS)
PHYSTGT:ID(STATE ENERGY COMPANY|ENERGY COMPANY)
INCIDENT:INSTRUMENTID(BOMB)
```

Matches between facts in this form and sentences are identified as belonging to one of three categories: **Full**, **Partial** or **NoMatch**. Each of these possibilities may be described as follows:

Full A fact fully matches a sentence if it mentions a filler for each field. For example, there is a Full match between the fact just shown and the following sentence since one of the possible fillers for each of the fields is mentioned in the sentence. `THE SHINING PATH CARRIED OUT BOMB ATTACKS AGAINST THE STATE ENERGY COMPANY.`

Partial A Partial match occurs when one of the fillers for at least two of the fields are listed in the sentence but there is not a Full match, i.e. at least one field is not mentioned. The following example demonstrates a Partial match, fillers for the `PERP` and `HUMTGT` fields are mentioned but not the `INCIDENT:INSTRUMENTID` field.

```
PERP(URBAN GUERRILLA COMMANDOS)
HUMTGT(JUAN CARLOS MERIOS| EMPLOYEE|ADMINISTRATIVE OFFICIAL)
INCIDENT:INSTRUMENTID(BULLET)
```

tor (`http://opennlp.sourceforge.net`) which had been retrained on a capitalised version of the Penn Treebank (Marcus et al., 1993).

[2] The process of converting potential answer keys into a regular expression includes escaping characters such as punctuation which are also metacharacters in the regex language used, allowing variable whitespace between tokens and concatenating each possible variation for a filler into a set of disjunctions.

FINALLY, ACCORDING TO A POLICE REPORT, ALLEGED URBAN GUERRILLA COMMANDOS THIS MORNING KILLED AN EMPLOYEE OF THE 1ST INFANTRY BRIGADE IN SAN SALVADOR.

NoMatch A fact does not match a sentence if the conditions for a Full or Partial match are not met. This occurs when there is no sentence in the corpus which contains more than one of the fact's fields and, in other words, the fillers of the fields which make up the fact are spread across separate sentences in the text. Although the fact's fields will appear in the text (since they are string slots) they must appear in the same sentence to be identified by SSA systems. The fact used as an example for the Partial match category would not match this sentence, despite containing the filler of the INCIDENT:INSTRUMENTID field: ACCORDING TO THE CORONER'S REPORT, MERIOS' BODY HAD FOUR BULLET WOUNDS.

To identify the proportion of facts which could be identified by an SSA system, each fact is compared against all the sentences in the corpus to identify the proportion which fully match at least one sentence. When this process has been completed the remaining facts are once again compared against the corpus to discover which of those partially match at least one sentence. We now go on to describe the results of experiments in which this process is applied to corpora from the MUC evaluations.

4. Experiments

A key decision in these experiments is the choice information in the templates to search for in the documents. The aim is to capture as much of the template's salient information as possible. The matching process (Section 3) is limited to considering string slots from templates. Some templates include string slots which are generally not instantiated because the information does not occur in the text and these are ignored for simplicity. In addition, some of the string slots contain similar information. For example, the MUC6 templates contain two slots which list ways of referring to the same person (PER_NAME and PER_ALIAS). These slots are combined and their possible fillers concatenated.

It is worth noting that this process of selecting the key information within templates and combining together field's values makes it more likely that the information will match the text (either fully or partially). Selectivity means that fewer pieces of information need to co-occur

within a sentence while combining slots requires that only one of them need to occur within a sentence to count as a match.

The aim of this work is not to make any claim about what constitutes a fact. In the context of this work a "fact" is considered to be any piece of information which can be identified within a text. The intention is to make use of the core information contained in a standard data set which is commonly used as a benchmark for IE systems. In two of these corpora (MUC4 and MUC7) the core information could be considered to be the description of events while in the MUC6 corpus the facts used for these experiments are more like descriptions of individuals.

4.1. Latin American Terrorism

The templates used for MUC4, such as the one shown in Figure 1, contained information describing terrorist incidents in Latin America (Sundheim, 1991). They included six string slots: (slot 6) `INCIDENT: INSTRUMENT ID`, the device used to carry out the act of terrorism; (9) `PERP: INDIVIDUAL ID`, person responsible for a terrorist incident; (10) `PERP: ORGANIZATION ID`, organisation responsible for a terrorist incident; (12) `PYHS TGT: ID`, any inanimate object that was the target of a terrorist act; (18) `HUG TGT: NAME`, any person who was the target, or became the victim of, an attack and (19) `HUM TGT: DESCRIPTION`, the title or role of a human target of a terrorist act or the general description of a unnamed human target. These slots contain some of the core information for each fact and were used to provide the fact definition for the experiment using the MUC4 corpus. Two pairs of slots contained similar information and were combined: slots 9 and 10 both describe the perpetrator of a terrorist act while slots 18 and 19 provide information about people who were the terrorist's targets. Consequently for these experiments the information examined consists of four fields which contain information about the identity of the perpetrator of a terrorist act (`Perp`), the target which may be either human (`Humtgt`) or physical (`Phystgt`) and the instrument used, such as a bomb (`Instrument`). The template shown in Figure 1 would be represented as follows:

```
Instrument(BOMB)
Perp(URBAN GUERRILLA GROUP|FMLN|
     FARABUNDO MARTI NATIONAL LIBERATION FRONT)
Humtgt(ROBERTO GARCIA ALVARADO|ATTORNEY GENERAL)
Phystgt(ARMORED VEHICLE)
```

4.2. Management Succession

The MUC6 corpus concerns management succession events. For example the sentence "Daniel Glass was named president of EMI Records Group, a unit of London's Thorn EMI PLC." describes an executive ("Daniel Glass") taking up a position ("president") within an organisation ("EMI Records Group"). The core information in the MUC6 templates is stored in a sub-template which lists (1) the person who is moving, (2) the organisation they are joining/leaving, (3) their post (job title) and (4) whether they are joining or leaving the organisation. The last piece of information is represented as a text conversion slot so the first three pieces of information were taken from the templates to form the facts for this corpus.[3] The fact shown above would be represented as follows:

```
Person('Daniel Glass'|'Glass')
Org('Thorn EMI PLC'|'EMI')
Post('president')
```

Alternative field fillers are identified by concatenating together fields in the MUC6 template which list various descriptions for entities.

4.3. Rocket Launches

The MUC7 task concerned information about rocket launches described in newswire reports. An example sentence from this corpus containing information of interest is the following: "In the early hours of Feb. 15, a new Chinese rocket took off from its launch pad in western Xinjiang province with a 205 million dollar satellite on board."

The core information in the MUC7 templates is stored in a sub-template which consists of fields containing details about the space vehicle, where it was launched and the payload being carried. A number of slots were concatenated to identify three keys pieces of information from the templates: `VEHICLE`, `LAUNCH_SITE` and `PAYLOAD`. The template containing information about the rocket launch described in the above sentence would be represented as follows:

```
VEHICLE('a new Chinese rocket'|'rocket')
PAYLOAD('a 205 million dollar satellite'|'satellite'|
        'Western Satellite')
LAUNCH_SITE('Xinjiang'|'China')
```

[3] In the MUC6 corpus the movement of the executive is often encoded in the text using a predicate-argument structure, e.g. "named" in the above example, although alternative structures may also be used, e.g. "Mr. Keller's resignation .....". It is difficult to identify these comprehensively in a reliable way and therefore attention is restricted to string slots.

4.4. Results: Fact Matches

Table I shows the result of the fact matching process described in Section 3 when applied to each of these corpora. The column marked "All" indicates the number of facts falling into each of the three categories and this is also expressed as a percentage. The columns marked "2", "3" and "4" show the proportion of facts consisting of two, three and four fields falling into each category. (The facts derived from the MUC6 and MUC7 corpora contain up to three fields and consequently there are no facts listed in the column marked "4" for these corpora.)

Table I. Counts of fact matches

Corpus	Match Type	Event fields All	2	3	4
MUC4	**Full**	718 (59.7%)	588	114	16
	Partial	226 (18.8%)	0	151	75
	NoMatch	256 (21.3%)	228	26	2
	Total	1200	816	291	93
MUC6	**Full**	336 (59.5%)	7	329	—
	Partial	225 (39.8%)	0	225	—
	NoMatch	4 (0.7%)	0	4	—
	Total	565	7	558	—
MUC7	**Full**	99 (63.1%)	73	26	—
	Partial	28 (17.8%)	0	28	—
	NoMatch	30 (19.1%)	25	5	—
	Total	157	98	59	—

It can be seen that the proportion of facts falling into the full match category is around 60% for all three corpora. This suggests that a SSA IE system could, at best, identify only hope to fully identify three fifth of the facts in these texts. Therefore it seems that the coverage of SSA systems is severely limited on these three corpora and that the approach is not sufficient to identify the information contained in these texts.

In each of the corpora around 40% of the facts fall into the partial and Nomatch categories. These facts cannot be fully identified by a SSA system. The distribution of facts across the partial and Nomatch categories is similar for the MUC4 and MUC7 corpora but differs for MUC6. In the MUC4 and MUC7 corpora both the partial and Nomatch categories contain around 20% of the facts. However, in the

MUC6 corpus 39.8% of the facts fell into the partial match category and only 0.7% were Nomatch'es. It is impossible for a SSA system to identify facts which fall into the Nomatch category, suggesting that these approaches may be more successful on the MUC6 corpus than the other two used in these experiments.

One reason for the low proportion of facts falling into the Nomatch category in the MUC6 corpus may be the relative simplicity of the facts derived from these text compared with the other two corpora. Management succession events in this corpus are often described within a comprehensive sentence, for example "QVC Network Inc., as expected, named Barry Diller its chairman and chief executive officer." Sentences which summarise the facts of interest occur less frequently in the other corpora. In addition, the MUC6 corpus contains a larger proportion of facts consisting of more than two fields than the other two corpora. Facts with two fields can either match the text fully or not at all while those with more fields can also participate in both Partial and Full matches. In MUC6 98.8% ($\frac{558}{565}$) of facts have at least three fields while this figure is just 32% ($\frac{384}{1200}$) for the MUC4 corpus and 37.6% ($\frac{59}{157}$) for MUC7.

4.5. Results: Field Matches

Table II shows an analysis of matches for individual fact fields. The pairs of figures in the main body of the table refer to the number of instances of the relevant field which are mentioned in the sentence matched by an event, identified by finding the sentence which matches the greatest number of fields for a particular fact, and the total number of instances of that field. The column headed "Full match" lists the facts which fully match the text and, as would be expected, all fields are matched. The columns marked "Partial match" and "NoMatch" lists the facts which fall into those categories. The "All matches" column shows the proportion of facts falling into either the Full or Partial Match categories and the total number of fields in the corpus. This figure is also expressed as a percentage.

It can be seen that there are differences between the percentage of matches both across the three corpora and for the various fields within each corpus. The highest proportion of matches is seen in the MUC6 corpus and it is likely that this is due to the fact that a higher proportion of facts in this corpus fall into the "Partial match" category compared with the other corpora and the relative simplicity of the facts in this corpus. The facts contained within the MUC4 corpus have the most complex structure, in terms of number of potential fields, and this may explain why the lowest matches are recorded for those texts.

Table II. Matches at field level

Corpus	Field	Match Full	Match Partial	Match NoMatch	All matches
MUC4	Perp	641/641	77/199	0/233	718/1073 (66.9%)
	Phystgt	304/304	32/199	0/85	336/588 (57.1%)
	Humtgt	496/496	51/163	0/181	547/840 (65.1%)
	Instrument	141/141	33/192	0/43	174/376 (46.3%)
	Total	1582/1582	193/753	0/542	1775/2887 (61.5%)
MUC6	Post	336/336	179/225	0/4	515/565 (91.2%)
	Org	329/329	99/225	0/4	428/558 (76.7%)
	Person	336/336	176/225	0/4	508/565 (89.9%)
	Total	1001/1001	454/675	0/12	1451/1688 (86.0%)
MUC7	LAUNCH_SITE	50/50	15/28	0/21	65/ 99 (65.66%)
	VEHICLE	82/82	21/28	0/22	103/132 (78.03%)
	PAYLOAD	92/92	20/28	0/32	112/152 (73.68%)
	Total	224/224	56/84	0/75	280/383 (73.1%)

Within each of the corpora it can be seen that there is some variation between individual fields in terms of the proportion of facts which match. In the MUC4 corpus the lowest results are observed for the Instrument field; less than half of the instances of this field participate in facts which match the text. Better performance is recorded for the Perp and Humtgt fields with around two thirds of instances participating in facts which match the text.

A reason for this difference is that fillers of the Perp and Humtgt or Phystgt fields often appear together in a sentence which summarises the incident and the filler of the Instrument field, which lists the weapon used, appears later in the text in a sentence which provides further detail. An example can be seen in the following pair of sentences from a MUC4 document which refer to an incident in which the Humtgt is "MARIA ELENA DIAZ PEREZ", Perp "10 PAID ASSASSINS" and the Instrument "SUBMACHINE GUN".

> MARIA ELENA DIAZ PEREZ, THIRD JUDGE OF PUBLIC ORDER, AND TWO OF HER BODYGUARDS FROM THE DAS [ADMINISTRATIVE DEPARTMENT OF SECURITY], WERE ASSASSINATED IN MEDELLIN TODAY BY A GROUP OF 10 PAID ASSASSINS IN TWO CARS. ...
> A TOTAL OF 55 9-MM SUBMACHINE GUN ROUNDS HIT THE LEFT SIDE OF THE CAR.

Results from the MUC6 corpus show that the `Post` and `Person` facts participate in matches more frequently than the `Org` field. This difference can also be explained by looking at the style in which the texts are written. In these documents management succession events are commonly introduced near the start of the newswire story and these descriptions almost invariably contain all three fact fields. For example, one story starts with the following sentence: "Washington Post Co. said Katharine Graham stepped down after 20 years as chairman, and will be succeeded by her son, Donald E. Graham, the company's chief executive officer." Later in the story further succession events may be mentioned but many of these use an anaphoric expression (e.g. "the company") rather than explicitly mention the name of the organisation in the event. For example, this sentence appears later in the same story: "Alan G. Spoon, 42, will succeed Mr. Graham as president of the company."

There is less difference between the percentage of the individual fields participating in a match in the MUC7 corpus. The documents which form this corpus tend to be less regular than the MUC4 and MUC6 documents (in which information of interest is often summarised at the start of the document and elaborated later). In these texts the facts to be identified tend to be distributed through the document and it is common to find sentences which contain two of the fact fields with another described separately. For example this pair of sentences from a MUC7 document shows the description of a rocket launch in which the `VEHICLE` ("Endeavour") and `LAUNCH_SITE` ("Kennedy Space Center") are mentioned in the first sentence and the `PAYLOAD` ("a $10 million NASA satellite") in the later one. "The shuttle Endeavour and a crew of six are to blast off Thursday at 4:18 a.m. EST from NASA's Kennedy Space Center. Midway through the mission, the crew plans to deploy a $10 million NASA satellite for nearly 48 hours of operations..."

In another example the `VEHICLE` ("Ariane 5") and `LAUNCH_SITE` ("Kourou, French Guiana") are mentioned in the first sentence and `PAYLOAD` ("four European Space Agency Cluster satellites") in the second.

> Kourou, French Guiana, June 4 (Bloomberg) – Ariane 5, a new and more powerful rocket developed by the pan-European Arianespace group, exploded within seconds of blastoff in a major setback to the world's leading commercial satellite launcher.
>
> The unmanned rocket, the most powerful yet built specifically for commercial payloads, was carrying four European Space Agency Cluster satellites, part of a $500 million project to study the interaction of the sun and the earth.

In summary, the style in which the documents are written has an effect on the facts which can be extracted from them using a SSA system. In some corpora, such as the ones used for MUC4 and MUC6, many of the facts are summarised in a single sentence at the start of the document. For these texts it would be feasible to extract certain pieces of information by examining single sentence contexts. For example, a SSA system could extract many of the relations between `Person` and `Post` in the MUC6 text, although it would be unable to identify many of the `Person` and `Org` relations. In other texts the information of interest is distributed in the documents in a less regular way. For example, MUC7 documents do not generally start with a summary of the rocket launches mentioned in the document and this information is normally distributed across the text. This suggests that SSA approaches may be more feasible for some extraction tasks than others and that the structure of documents from which information is being extracted is important.

4.6. Alternative Analysis

It has already been mentioned that the approach described here estimates an upper bound on the proportion of facts which are described within single sentences. Stevenson (2004) reports an alternative approach which places a more accurate bound on this figure, but required additional data and could only be applied to the MUC6 corpus. This approach made use of an alternative version of the MUC6 corpus, produced by Soderland (1999), in which only facts described within a single sentence were annotated. This set of facts was compared with the ones extracted from the MUC6 templates (which include all facts mentioned in the documents). Each fact derived from the templates was identified as being either a full match, partial match or nomatch, with these categories being analogous to the definitions used here: a full match was said to occur when a fact derived from the MUC6 template was also listed in Soderland's version of the corpus, a partial match when at least two of the fields match for facts in both corpora and Nomatch when a fact in the MUC6 corpus is not mentioned in Soderland's version. This approach is more accurate than the one used here because a match (full or partial) occurs when a sentence genuinely mentions a fact, not just when the string slots occur together. These experiments used the same fields to define a fact as used here (post, organisation and person). Stevenson (2004) reported that 40.6% of the facts fell into the Full match category, 39.1% were partial matches and the remaining 20.3% Nomatches.

The number of facts categorised as full matches is substantially less than the one reported here (59.5%). We do not have access to corpora annotated with events at the sentence level which are necessary to carry out this analysis for the MUC4 and MUC7 corpora so it is not possible to generate comparable results for these data sets. It may also be problematic to try to infer too much about how these results may effect other corpora given that the fact structure is less complex in MUC6. However, the difference in these results suggests that the true proportions may be substantially lower than the figures reported in this paper.

5. Combining Facts Across Sentences

The experiments described so far show that it is not possible to identify a substantial proportion of facts within a document by only examining each sentence in isolation. This naturally raises the question of how these facts can be identified. Analysis of the documents used for these experiments show that various linguistic devices are used to connect the parts of a fact description across sentences.

The most straightforward of these is when an anaphoric expression is used to refer to one of the fact's fields. For example, this pair of sentences appear in the MUC6 corpus: "Wall Street was hoping for stronger outside management to help Figgie. Instead, the company named a director, 66-year-old Walter M. Vannoy, who has been on the board since 1981." The second sentence describes the promotion of Walter M. Vannoy to the position of director in a company called Figgie. However, the name of the company is not mentioned directly but is referred to by an anaphoric expression. In these cases the fact could be considered to be described entirely within one sentence but with some fields being referred to indirectly. We refer to these cases as single sentence facts containing anaphoric references.

In more complex cases the fact description is genuinely spread across more than one sentence with the various parts of the description being linked by anaphoric expressions or alternative descriptions. For example, Section 4.5 shows two sentences from the MUC7 corpus describing the launch of the "Ariane 5" rocket. The first sentence mentions the vehicle and launch site while the second contains details of its payload. The two sentences are connected through the phrase "unmanned rocket" but neither sentence contains all three fields which form this fact, even if anaphoric expressions are resolved. Another example, this one from the MUC6 corpus, is shown in Section 1 where the name of the organisation ("Pace America Group Inc.") is mentioned in the first

sentence and the name of two executives leaving that company and their positions in the second. Although the sentences are connected by the coreference chain connecting "The executives" and "two top executives" neither contains all three of the fields which form the fact, either directly or indirectly via coreference. These cases are referred to as connected multiple sentence facts. For cases such as these some inference will be required to combine together all the parts of the fact description.

In the cases discussed so far the various parts of the fact description are connected via some referential relationship in the text. However, in other cases there may be no direct connection between the sentences describing the fact. These facts can only be identified using a deeper understanding of the text such as discourse analysis or the application of world knowledge. For example, the two sentences from a MUC4 document shown on page 12 describe an assassination. The main description of the incident is listed in the first sentence and the instrument ("SUBMACHINE GUN") in the second. These pieces of information can only be combined to form a complete fact with knowledge that the main topic of this document is the assassination and the second sentence provides detail about it. There is no direct connection between the two sentences. (Note that the noun phrase "THE TWO CARS" in the first sentence is not the antecedent of "THE CAR" in the second sentence.) This is an example of a situation in which the various parts of the fact are described in text without being directly connected. There are other cases in which information is not mentioned in the text but has to be inferred using world knowledge. For example, the following two sentences are taken from the MUC6 corpus: "David J. Bronczek, vice president and general manager of Federal Express Canada Ltd., was named senior vice president, Europe, Africa and Mediterranean, at this air-express concern. Mr. Bronczek succeeds Kenneth Newell, 55, who was named to the new post of senior vice president, retail service operations." This text describes two movements of position for Kenneth Newell: leaving the position of vice president and moving to the position of senior vice president. However, the fact that Newell is leaving a position can only recognised with knowledge that when one executive replaces another then that executive must leave their current position. Cases such as these are referred to as unconnected multiple sentence facts.

Single sentence facts containing anaphoric expressions are likely to be the most straight forward to identify automatically since they do not require the combination of information in separate sentences. However, multiple sentence facts, both connected and unconnected, require inference and, possibly, the application of world knowledge to be recognised.

The remainder of this Section describes an experiment which quantifies the proportion of single sentence facts containing anaphoric expressions in the MUC6 corpus.

5.1. Experiment

In order to estimate the proportion of single sentence facts containing anaphoric expressions we require a corpus for which the facts have been identified (such as those used for the experiments in Section 4) and some method for resolving anaphoric expressions in those texts. Any automatic system for anaphora resolution will make errors so we prefer to make use of manual annotation. Fortunately, portions of the MUC6 and MUC7 corpora were manually annotated with coreference chains as part of the evaluation and are ideal for this purpose.[4] However, only a small portion of each corpus was annotated with this information (presumably because of the cost of annotation). The MUC6 corpus contains 20 texts which are annotated with coreference information and have facts associated with them. These texts contain a total of 97 facts. The MUC7 corpus does not contain any appropriate documents since none of those which are annotated with coreference information contain any facts.[5] The 20 texts from the MUC6 corpus were used for the experiments described in this Section and are referred to as the "coreference corpus".

Texts in this corpus are annotated with coreference chains. Each coreferential expression is labelled with a unique identifier and the identifier of its immediate antecedent. Figure 2 shows two sentences from the MUC6 corpus annotated with coreference information (slightly simplified for clarity).[6] The second sentence is a single sentence fact containing an anaphoric expression. It describes the fact that Alan G. Spoon will become president of Washington Post Co., although the name of the company is referred to indirectly using a corefential expression ("the company"). The annotation shows that the referent of this expression is one which has been labelled with the identifier `11` ("the company" in the first sentence) and that its reference is the one labelled `2`: "Washington Post Co."

[4] In addition to the IE task the MUC6 and MUC7 evaluations included a number of other language processing tasks, including coreference resolution.

[5] The corpora used for the various MUC evaluations contain a mixture of relevant documents (which contain facts) and non-relevant documents (which do not).

[6] In the annotation format used for this corpus anaphoric expressions and their antecedents are enclosed in `<COREF> ... </COREF>` SGML tags. The unique identifier of each expression is denoted by the `ID` attribute and the antecedent of a anaphoric expression by `REF`.

```
<COREF ID="2">Washington Post Co.</COREF> said <COREF ID="4"
REF="5">Katharine Graham</COREF> stepped down after 20 years
as <COREF ID="6">chairman</COREF>, and will be succeeded by
<COREF ID="8" REF="0"><COREF ID="7" REF="4">her</COREF> son,
<COREF ID="9" REF="8">Donald E. Graham</COREF>, <COREF ID="10"
REF="8"><COREF ID="11" REF="2">the company</COREF>'s chief
executive officer</COREF>.</COREF>
<COREF ID="32">Alan G. Spoon, 42,</COREF> will succeed
<COREF ID="30" REF="29">Mr. Graham</COREF> as <COREF
ID="31" REF="32">president of <COREF ID="33" REF="11">the
company</COREF></COREF>.
```

Figure 2. Example text from the coreference corpus

This data was used to carry out an experiment to determine the proportion of single sentence facts containing anaphoric expressions in the coreference corpus. The experiment was based on the matching process described in Section 3. However, rather than requiring fillers of the fields which constitute a fact to appear together within a sentence, we also consider them to co-occur if a possible filler is one of the possible antecedents of an expression which occurs within the sentence. The filler can occur anywhere in the text before the expression and does not need to be the immediate antecedent.

For example, using the matching process described in Section 3 (which does not include the antecedents of anaphoric expressions in possible matches) a fact with the following fields `person(Alan G. Spoon)`, `post(president)` and `org(Washington Post Co.)` would match the second sentence in Figure 2 only partially. However, when the antecedents of coreferential expressions are also allowed to participate in matches this fact would fully match that sentence.

This procedure, like the one described in Section 3, places an upper bound on the number of facts which could be matched. Annotation of the MUC data represents perfect anaphora resolution and it is unlikely that this result could be repeated in an actual system.

The experiments included two levels of anaphora resolution: all and pronominal. In the first the antecedents of all anaphoric expressions are examined to identify matches. When anaphora resolution is restricted to anaphora only the antecedents of pronominal anaphora expressions participate in matches. Pronominal anaphora is examined in isolation because it is the most common form of anaphora (Mitkov, 2003, p. 268) and this experiment is designed to determine how much can be gained when it is used alone. These two approaches are compared with the case when no anaphora resolution is used (referred to as "None") which is identical to the matching process outlined in Section 3.

5.2. Results

The results of this experiment are shown in Table III. For each level of anaphora resolution (all, pronominal and none) the proportion of facts falling into the full, partial and nomatch categories is shown. Each of the 97 facts in the coreference corpus consisted of three fields so the results are not broken down by number of field (unlike those reported in Table I).

A first observation is that the proportion of facts falling into the full match category when no coreference resolution is carried out is around 54%. This figure is lower that the one recorded when all texts in the MUC6 corpus were included in the analysis (see Table I). This shows that there is variation in the proportion of facts which are expressed within a single sentence and may also indicate that the events contains in these particular corpora are more distributed than the rest of the MUC6 corpus.

When all anaphoric expressions are resolved almost 20% more facts are fully matched. However, over a quarter of the facts are still only partially matched and these must be multiple sentence facts (either connected or unconnected). The application of pronominal anaphora resolution allows 6% more facts to be fully matched than when no anaphora resolution was applied. This demonstrates that, while the resolution of pronominal anaphora is useful, there is a significant benefit from resolving as wide a range of anaphoric expressions as possible.

Table III. Count of fact matches on MUC6 corpus with various levels of anaphora resolution

Match	Anaphora Resolution		
	None	Pronominal	All
Full	52 (53.6%)	58 (59.8%)	71 (73.2%)
Partial	43 (44.33%)	38 (39.2%)	26 (26.8%)
NoMatch	2 (2.06%)	1 (1%)	0 (0%)

The results of a field by field analysis for these experiments is shown in Table IV, which uses a similar format to Table II. It can be seen that matches for the `Post` and `Person` fields remain consistent for the various levels of anaphora resolution; matches for the `Post` field vary by a little more than 5% and `Person` not at all. However, a far larger variation, over 20%, is observed for the `Org` field. This is consistent with the analysis in Section 4.5 which showed that the `Org` field of

the fact was often linked with the other fields through a coreferential expression.

Table IV. Count of fact matches on MUC6 corpus with various levels of anaphora resolution

Anaphora Resolution	field	full	partial	nomatch	TOTAL
none	`Post`	52/52	33/43	0/2	85/97 (85.57%)
	`Person`	32/52	34/43	0/2	84/97 (86.60%)
	`Org`	52/52	21/43	0/2	73/97 (75.26%)
pronominal	`Post`	58/58	27/38	0/1	85/97 (87.63%)
	`Person`	58/58	26/38	0/1	84/97 (86.60%)
	`Org`	58/58	21/38	0/1	79/97 (81.44%)
all	`Post`	71/71	17/26	0/0	88/97 (90.72%)
	`Person`	71/71	13/26	0/0	84/97 (86.60%)
	`Org`	71/71	22/26	0/0	93/97 (95.88%)

These results show that the use of anaphora resolution leads to a substantial increase in the proportion of facts which can be identified by SSA systems. The antecedents of a full range of anaphoric expressions need to be identified to realise this benefit. However, field analysis shows that this may benefit some pieces of information more than others and care should be taken to ensure that anaphora resolution will be of benefit for a particular extraction task. For example, an IE system which aims to identify relations between `Person` and `Post` in the MUC6 corpus will not gain substantially from the use of anaphora resolution but this would be highly beneficial for the `Person-Org` relation.

Even when anaphoric expressions are resolved a significant proportion of the facts in the MUC6 corpus could not be fully identified by a SSA system. These facts require inference across the information contained in various sentences to be identified, possibly using discourse analysis and world knowledge. An IE system which aims for comprehensive identification of facts in text must then make use of these techniques and cannot rely on simpler approaches.

Unfortunately the results reported in this Section are limited by the fact that there is only a small amount of data which is suitable for these experiments available. However, they do indicate that anaphora resolution will help in the process of fact identification and that the resolution must be carried out over as wide a range of anaphoric expressions as possible.

6. Related Work

Hirschman (1992) carried out an analysis of the difficulty of the MUC4 evaluation set. She categorised each document as requiring a single template or multiple templates to be filled. In addition the information which filled these templates was classed as being found in either a single sentence or multiple sentences. It was discovered that document requiring the filling of more than one template were easier when the information for each template was contained within a single sentence than when it was spread across multiple sentences. However, an unexpected result was that documents which require a single template to be filled where the information was contained within a single sentence were actually more difficult than those where the information was spread across multiple sentences. Hirschman attributed this to the fact that these documents were mainly comprised of irrelevant information and that the process of identifying this overshadowed the difficulty of combining information across sentences. It was also found that the performance of different systems across documents was very consistent which implied that some texts are more difficult to understand than others.

Bagga and Biermann (1997) developed techniques for comparing the difficulty of IE tasks by assigning a "domain number" which represented the complexity of the facts being extracted. They found that the MUC6 evaluation task was easier than the one used for the fifth MUC (international joint ventures) but harder than the one used for MUC4. However, Bagga and Biermann's technique did not take into account the distribution of the facts in text.

The results presented by Hirschman and the analysis presented here shows that the description of different parts of facts can be distributed through a text. Huttunen et al. (2002) also demonstrated that facts which are described in this way are more difficult to identify.

7. Summary and Implications

The experiments described here show that a substantial proportion of facts in three commonly used IE evaluation corpora are not expressed within a single sentence and, therefore, cannot be identified by SSA systems. These experiments used a variety of domains and text types including newswire text, broadcast news and transcribed speech. Further experiments using anaphora resolution show that, while it is beneficial to IE systems, a deeper level of understating of text is required to identify all facts contained within documents. The exact proportion of

facts which cannot be expressed within a single sentence is perhaps not particularly significant in itself, and will depend upon the particular documents and facts being extracted from them. The procedures outlined here place upper bounds on the proportion of facts which are expressed within a single sentence and the true maximum performance for SSA systems may be even lower.

These results have implications for the evaluation of IE algorithms. Many recent systems have been evaluated in terms of their ability to extract facts which are expressed within single sentences, for example (Yangarber et al., 2000; Soderland, 1999; Chieu and Ng, 2002; Zelenko et al., 2003; Stevenson and Greenwood, 2005), and the analysis reported here demonstrates that the results for these approaches are are likely to be significantly lower if those systems attempted to extract all facts. Results from SSA systems should be reinterpreted in light of this information.

These results should also be taken into account in the implementation of IE systems. Approaches which fail to consider facts whose description is spread across several sentences are unable to fully identify up to 60% of the facts in the three corpora analysed for these experiments. So, applications which require a comprehensive set of facts to be extracted from a document, particularly if those facts are not simple binary relations, must ensure that their systems can identify those expressed across multiple sentences. In addition, IE systems which aim to identify all facts in text must make use of relatively deep analysis of text, including modelling of the discourse and use of world knowledge.

The experiments reported here also show that some facts (for example the relation between `Person` and `Post` in the MUC6 corpus) are more likely to be described within a single sentence than others. SSA systems may be an appropriate technique for identifying these facts. However, it is important for text to be analysed to determine whether the facts of interest are stated this way before an approach which considers each sentence in isolation is chosen.

Acknowledgements

This work was carried out as part of the RESuLT project, funded by the UK EPSRC (GR/T06391). I am grateful to Mark Hepple, Mark Greenwood, David Martinez and Paul Clough for providing feedback on earlier versions of this paper. Any mistakes are my own.

References

Bagga, A. and A. Biermann: 1997, 'Analyzing the Complexity of a Domain with Respect to an Information Extraction Task'. In: *Proceedings of the Tenth International Conference on Research on Computational Linguistics (ROCLING-X)*. Taipei, Taiwan, pp. 174–194.

Chieu, H. and H. Ng: 2002, 'A Maximum Entroy Approach to Information Extraction from Semi-structured and Free Text'. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence (AAAI-02)*. Edmonton, Canada, pp. 768–791.

Culotta, A. and J. Sorensen: 2004, 'Dependency Tree Kernels for Relation Extraction'. In: *42nd Annual Meeting of the Association for Computational Linguistics*. Barcelona, Spain, pp. 423–429.

Grishman, R.: 2003, 'Information Extraction'. In: R. Mitkov (ed.): *The Oxford Handbook of Computational Linguistics*. Oxford University Press, pp. 545–559.

Grover, C., C. Matheson, A. Mikheev, and M. Moens: 2000, 'LT TTT - A Flexible Tokenisation Tool'. In: *Proceedings of Second International Conference on Language Resources and Evaluation (LREC 2000)*. Athens, Greece.

Hirschman, L.: 1992, 'An Adjunct Test for Discourse Processing in MUC-4'. In: *Proceedings of the Fourth Message Understanding Conference (MUC-4)*. San Francisco, CA, pp. 67–77.

Huttunen, S., R. Yangarber, and R. Grishman: 2002, 'Complexity of Event Structures in IE Scenarios'. In: *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*. Taipei, Taiwan, pp. 376–382.

Marcus, M., B. Santorini, and M. Marcinkiewicz: 1993, 'Building a Large Annotated Corpus of English: The Penn Tree Bank'. *Computational Linguistics* **19**(2), 313–330.

Mitkov, R.: 2003, 'Anaphora Resolution'. In: R. Mitkov (ed.): *The Oxford Handbook of Computational Linguistics*. Oxford University Press, pp. 266–283.

Sekine, S.: 2006, 'On-Demand Information Extraction'. In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Sydney, Australia, pp. 731–738.

Soderland, S.: 1999, 'Learning Information Extraction Rules for Semi-structured and free text'. *Machine Learning* **31**(1-3), 233–272.

Stevenson, M.: 2004, 'Information Extraction from Single and Mutliple Sentences'. In: *Proceedings of the Twentieth International Conference on Computational Linguistics (COLING-02)*. Geneva, Switzerland, pp. 875–881.

Stevenson, M. and M. Greenwood: 2005, 'A Semantic Approach to IE Pattern Induction'. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbour, MI, pp. 379–386.

Sundheim, B.: 1991, 'Overview of the Third Message Understanding Evaluation and Conference'. In: *Proceedings of the Third Message Understanding Conference (MUC-3)*. San Diego, CA, pp. 3–16.

Yangarber, R., R. Grishman, P. Tapanainen, and S. Huttunen: 2000, 'Automatic Acquisition of Domain Knowledge for Information Extraction'. In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. Saarbrücken, Germany, pp. 940–946.

Zelenko, D., C. Aone, and A. Richardella: 2003, 'Kernel Methods for Relation Extraction'. *Journal of Machine Learning Research* **3**, 1083–1106.