Generation of Repeated References to Discourse Entities

Anja Belz
Natural Language Technology Group
University of Brighton
A.S.Belz@brighton.ac.uk

Sebastian Varges
Information and Communication Technology
University of Trento
varges@dit.unitn.it

Abstract

Generation of Referring Expressions is a thriving subfield of Natural Language Generation which has traditionally focused on the task of selecting a set of attributes that unambiguously identify a given referent. In this paper, we address the complementary problem of generating repeated, potentially different referential expressions that refer to the same entity in the context of a piece of discourse longer than a sentence. We describe a corpus of short encyclopaedic texts we have compiled and annotated for reference to the main subject of the text, and report results for our experiments in which we set human subjects and automatic methods the task of selecting a referential expression from a wide range of choices in a full-text context. We find that our human subjects agree on choice of expression to a considerable degree, with three identical expressions selected in 50% of cases. We tested automatic selection strategies based on most frequent choice heuristics, involving different combinations of information about syntactic MSR type and domain type. We find that more information generally produces better results, achieving a best overall test set accuracy of 53.9% when both syntactic MSR type and domain type are known.

1 Introduction

Generation of Referring Expressions (GRE) is one of the most lively and thriving subfields of Natural Language Generation (NLG). GRE has traditionally addressed the following question:

> [G]iven a symbol corresponding to an intended referent, how do we work out the semantic content of a referring expression that uniquely identifies the entity in question? (Bohnet and Dale, 2005, p. 1004)

This view of GRE is mainly concerned with ruling out 'distractors' to achieve unique identification of the target referent. Our research is concerned with a complementary question: given an intended referent and a discourse context, how do we generate appropriate referential expressions (REs) to refer to the referent at different points in the discourse? While existing GRE research has taken discourse context into account to some extent (see Section 2), the question why people choose different REs in different contexts has not really been addressed:

> Not only do different people use different referring expressions for the same object, but the same person may use different expressions for the same object on different occasions. Although this may seem like a rather unsurprising observation, it has never, as far as we are aware, been taken into account in the development of any algorithm for generation of referring expressions. (Viethen and Dale, 2006, p. 119)

Selection of a particular RE in a particular context is likely to be affected by a range of factors in addition to discourse-familiarity and unique identification. In our research we ultimately aim to (i) investigate the factors that influence choice of RE in context, (ii) determine what information is needed for a GRE module to be able to generate appropriate REs in context, and (iii) develop reliable methods for automatically generating REs in context.

Our basic approach is to annotate occurrences of MSR in naturally occurring texts, analyse the texts in various ways, and obtain multiple, human-produced alternatives to the REs in the texts. The results are used to inform the design of automatic methods for RE selection. The success of such methods can in turn be evaluated in terms of similarity of output REs with the human-produced REs.

In our current work we are focusing on a text type that has a single, easily identifiable main subject for which we can therefore expect to find a range of different REs: encyclopaedic entries. In this paper, we describe a corpus of such texts we have compiled and annotated (Section 3), and report first insights from our analysis of the corpus data (Section 4). We further report the results of an experiment where subjects selected REs in context (Section 5), and establish baseline results for automatic methods of selection (Section 6).

2 Related Research

The most classical form of GRE algorithm takes into account two main factors in selecting expressions: unique identification (of the intended referent from a set including possible distractors), and brevity (Dale, 1989; Reiter and Dale, 1992). Most GRE research focuses on definite, non-first mentions of the target referent. The most influential of these algorithms, the 'incremental algorithm' (IA) (Dale and Reiter, 1995), originally just selected attributive properties, but a range of extensions have been reported. Siddharthan and Copestake's algorithm (2004) is able to identify attributes that are particularly discriminating given the entities in the contrast set, and van Deemter's SET algorithm can generate REs to sets of entities (van Deemter, 2002).

Krahmer and Theune (2002) moved away from unique identification, also taking discourse context into account: they replaced the requirement that the intended referent be the *only* entity that matches the RE, to the requirement that it be the *most salient* in a given context. Several versions of centering theory have been used as a basis for pronominalisation algorithms (Dale, 1992; McCoy and Strube, 1999; Henschel et al., 2000). Jordan (2002) highlighted a factor other than salience that influences choice of RE: she found a large proportion of overspecified redescriptions in the Coconut corpus of dialogues and showed that some dialogue states and communicative goals make overspecific REs more likely.

Among the few corpora of texts within which REs have been annotated in some way (as opposed to corpora of annotated REs such as those created by van Deemter et al. (2006)) are the GNOME, Coconut and Maptask corpora. In the GNOME Corpus (Poesio, 2000; Poesio, 2004) different types of discourse and semantic information are annotated, including reference and semantic attributes. The corpus annotation was e.g. used to train a decision tree learner for NP modifier generation (Cheng et al., 2001).

The RE annotations in the Coconut corpus represent information at the discourse level (reference and attributes used) and at the utterance level (information about dialogue state). The 400 REs and annotations in the corpus were used to train an RE generation module (Jordan and Walker, 2000). Gupta and Stent (2005) annotated both the Maptask and Coconut corpora for POS-tags, NPs, referent of NPs, and knowledge representations for each speaker which included values for different attributes for potential referents.

While context has been taken into account to some extent in existing research on generation of REs, our goal is to model a range of contextual factors and the interactions between them. Our corpus creation work provides — for the first time, as far as we are aware — a resource that includes multiple human-selected REs for the same referent in the same place in a discourse. In contrast to the resources cited above, our corpus is a collection of naturally occurring texts. It is also somewhat larger, containing approximately 8,000 REs in total.

3 The Corpus

We created a corpus of short encyclopaedic texts by collecting just over 1,000 introductory sections from Wikipedia entries for cities, countries, rivers and people. An introductory section was defined as the part of the entry preceding the table of contents (we only used entries with tables of contents). We removed Wikipedia mark-up, images, HTML tags etc. from the entries to yield text-only versions. These were then annotated for references to the subject of the entry by five annotators, and the annotations double-checked by the first author. Annotators managed to do between 5 and 10 texts per hour. The inter-annotator agreement was 86%, as checked on a randomly selected 20-text subset of the corpus for which we had annotations by all five annotators (these annotations were not double-checked). The final corpus consists of 1,078 texts in four subdomains: rivers (83 texts), cities (248 texts), countries (255 texts) and people (492 texts).

3.1 Types of referential expression annotated

We annotated three broad categories of main subject referential expressions (MSREs) in our corpus[1] — subjects, objects and possessives. These are rel-

---

[1] In our terminology and view of grammar in this section we rely heavily on Huddleston and Pullum (2002).

atively straightforward to identify, and account for virtually all cases of main subject reference (MSR) in our texts. Annotators were asked to identify subject, object and possessive NPs and decide whether or not they refer to the main subject of the text. The three MSR types were defined as follows (NPs that we annotated are underlined):

I Subject MSREs: referring subject NPs, including pronouns and special cases of VP coordination where the same MSRE is the subject of the coordinated VPs, e.g:

1. *He was proclaimed dictator for life.*
2. *Alexander Graham Bell (March 3, 1847 - August 2, 1922) was a Scottish scientist and inventor who emigrated to Canada.*
3. *Most Indian and Bangladeshi rivers bear female names, but this one has a rare male name.*
4. *"The Eagle" was born in Carman, Manitoba and ___ grew up playing hockey.*

II Object MSREs: referring direct or indirect objects of VPs and prepositional phrases; e.g.:

1. *People from the city of São Paulo are called paulistanos.*
2. *His biological finds led him to study the transmutation of species.*

III Possessive MSREs: genitive NPs including genitive forms of pronouns, but excluding genitives that are the subject of a gerund-participial[2]:

1. *Its estimated length is 4,909 km.*
2. *The country's culture, heavily influenced by neighbours, is based on a unique form of Buddhism intertwined with local elements.*
3. *Vatican City is a landlocked sovereign city-state whose territory consists of a walled enclave within the city of Rome.*

3.2 Comments on annotation scheme

We interpret relative pronouns in a particular type of relative clause as anaphorically referential (I(2) and III(3) above): the type that Huddleston and Pullum call *supplementary relative clauses* (as opposed to integrated relative clauses). The main difference in meaning between the two types of relative clause is that in supplementary ones, the relative clause can be dropped without affecting the meaning of the

clause containing it. From the point of view of generation, the meaning could be equally expressed in two independent sentences or in two clauses one of which is a relative clause. The single-sentence construction is very common in the People subdomain of our corpus. One example is shown in (1) below, with the semantically equivalent two-sentence alternative shown in (2):

(1) *Hristo Stoichkov is a football manager and former striker who was a member of the Bulgaria national team that finished fourth at the 1994 FIFA World Cup.*

(2) *Hristo Stoichkov is a football manager and former striker. He was a member of the Bulgaria national team that finished fourth at the 1994 FIFA World Cup.*

We also annotated 'non-realised' MSREs in a restricted set of cases of VP coordination where an MSRE is the subject of the coordinated VPs. Consider the following example, where the subclausal coordination in (3) is semantically equivalent to the clausal coordination in (4):

(3) *He stated the first version of the Law of conservation of mass, ___ introduced the Metric system, and ___ helped to reform chemical nomenclature.*

(4) *He stated the first version of the Law of conservation of mass, he introduced the Metric system, and he helped to reform chemical nomenclature.*

According to Huddleston and Pullum (p. 1280), utterances as in (3) can be thought of as a reduction of longer forms as in (4), even though the former are not syntactically derived by ellipsis from the latter. Our reason for annotating the approximate place where the subject NP would be if it were realised (the gap-like underscores above) is that from a generation perspective there is a choice to be made about whether to realise the subject NP or not. Note that because we only included cases where subclausal coordination is at the level of VPs, these are all cases where only the subject NP is 'missing'[3].

Apart from titles and anything in quotations we included all NPs in our analysis. There are other forms of MSR that we could have included in our analysis, but decided against, because annotation simply proved too difficult: MSRs that are true gaps

---

[2]E.g. *His early career was marred by *his being involved in a variety of social and revolutionary causes*.

[3]E.g. we would not annotate a non-realised MSRE e.g. in *She wrote books for children and books for adults.*

and ellipses, adjective and noun modifiers, and implicit or anaphorically derivable references (other than those mentioned above).

4 Examining the Evidence

During the annotation process, the annotators found that the question 'does this expression refer to the main subject of this entry' was not always straightforward to answer. Consider the following passage:

(5) *A troop of Siberian Cossacks from Omsk founded the fort Zailiysky in 1854 at the foot of the Tian Shan mountain range, and renamed <u>it</u> one year later to Vernyj, a name that remained until 1921. In 1921, the name Alma-Ata ("father-apple") was created by the Bolsheviks. In a devastating earthquake in 1911, almost the only large building that remained standing was the Russian Orthodox cathedral. In the 1920s, after the completion of the Turkestan-Siberia Railway, <u>Alma-Ata</u>, as <u>it</u> was then known, became a major stopping point along the track. In 1929, <u>Almaty</u> became the capital of the Kazakh SSR.*

The actual MSREs (*Fort Zailiysky*, *Alma-Ata*, *it*, *Almaty*) are underlined, but there are a range of other terms that could be used to refer to the main subject (*father-apple*, *Vernyi*, *the capital of the Kazakh SSR*). There are three main issues. The first is metalinguistic use[4] of potential REs (as in *the name Alma-Ata* above) which did not cause major difficulties. Another issue is lexical ambiguity, e.g. an occurrence of *Australia* could refer to the continent or the country, and *Dubai* could refer to the city or the emirate. However, by far the more difficult issue arises where, if there are two referents, they cannot be said to be entirely distinct. Consider the following examples:

(6) *<u>The Indus system</u> is largely fed by the snows and glaciers of the Karakoram, Hindu Kush and Himalayan ranges. The Shyok, Shigar and Gilgit streams carry glacieral waters into <u>the main river</u>.*

(7) *<u>Aruba's</u> climate has helped tourism as visitors to <u>the island</u> can reliably expect warm, sunny weather.*

In (6) if one were to say that *the main river* and *the Indus system* had two distinguishable referents, the relation between them would clearly be one of part and whole. In (7), it could be argued that there

---

[4] "[T]he case where we cite a linguistic expression in order to say something about it qua linguistic expression." (Huddleston and Pullum, 2002, p. 401).

are two referents (the country Aruba and the geological formation that it occupies), but this is not entirely satisfactory. One of the aspects of a country is its geographical dimension, so *the island* could be said to refer to that aspect of Aruba.

These issues are simpler in the People subdomain (and this is the reason why we decided to include more people entries in the corpus): at least it is fairly clear when and where people begin and end, but there are still many 'partial' references, e.g. *the young man* in the following sentence:

(8) *<u>His</u> aptitude was recognized by <u>his</u> college headmaster, who recommended that <u>the young man</u> apply for the École Normale Supérieure.*

It is clearly not entirely a matter of deciding whether two REs refer to two distinct referents or to the same referent, but there appear to be a whole range of intermediate cases where referents are neither identical nor entirely distinct. Most REs refer to one or more aspects of a referent more strongly than the others. E.g. *the island* refers most strongly to the geographical aspect of Aruba, *the democracy* to its political aspect, and so on. However, there also appear to be default REs that are neutral with regard to these different aspects, e.g. *Aruba* in the current example.

From the point of view of the generation process, the fact that some potential REs refer to one or more aspects of the intended referent more strongly than others is important, because it is one of the reasons why different REs are chosen in different contexts, and this is an issue largely orthogonal to discourse-familiarity, addressee-familiarity and whether the intended referent as a whole is in focus or not.

Such matters of aspectual focus are likely to interact with other discourse-level and contextual factors that may influence choice of RE in a repeated reference, such as salience, discourse focus and structure, distance from last mention, presence of potential distractors, and text genre.

5 Human Choice of MSR

We had two reasons for conducting an experiment with human subjects as described below. For one, we wanted to get an idea of the degree to which RE choice followed patterns that we could hope to replicate with automatic methods. If our subjects agreed substantially, then this would seem likely. The other reason was that we needed a reference test set with multiple REs (in addition to the corpus
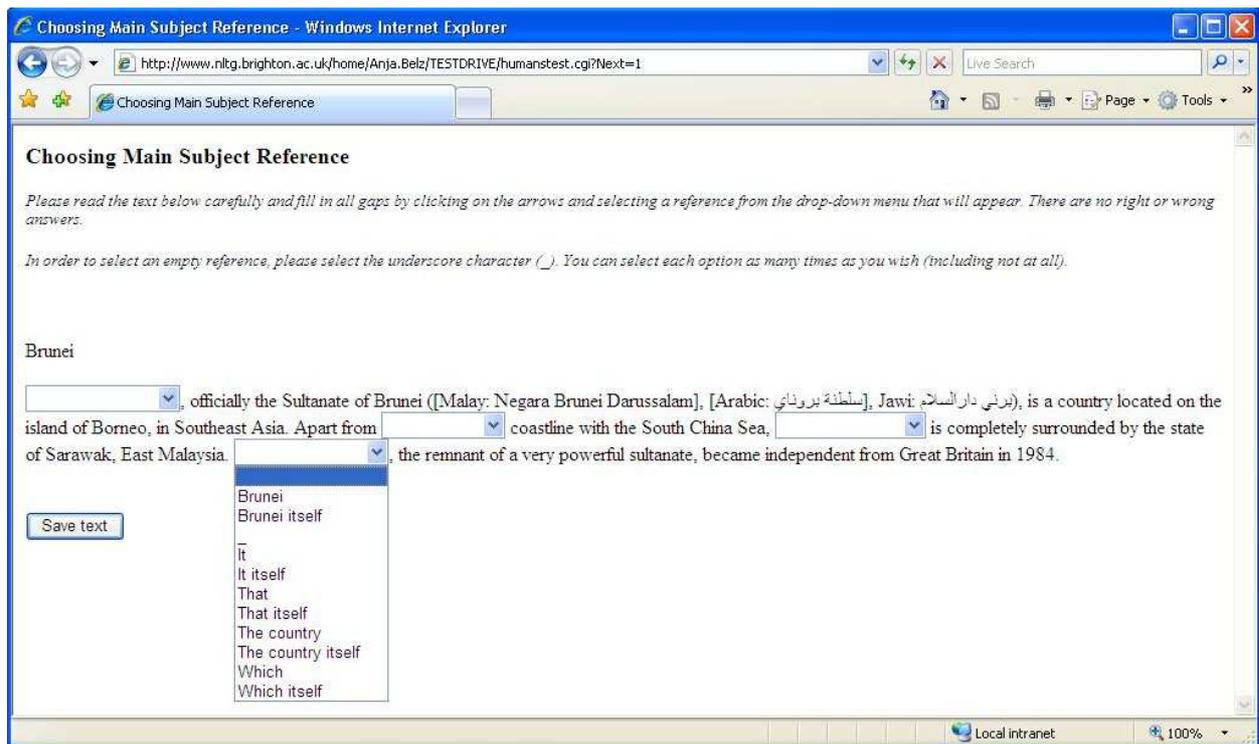
Figure 1: Screen shot of Choosing MSR Experiment.

texts) for each MSR to evaluate our automatic methods against, as is standard e.g. in MT and document summarisation.

We randomly selected a 10% subset of our corpus as our test set, ensuring that there were an equal number of texts from each subdomain and annotator. We then conducted an experiment in which we deleted all annotated MSREs and asked subjects to select an RE from a list of possible REs. Subjects were asked to do at least three texts each, over the web in a set-up as shown in Figure 1. The list of possible REs was automatically generated from the REs that actually occurred in each text, also using some additional generic rules (e.g. adding REs based on category nouns such as *the country* in the screen shot). We did not monitor who did the experiment, but asked members of the Corpora mailing list, colleagues and friends to participate anonymously. Approximately 80 different subjects did the experiment. Texts were randomly selected for presentation to subjects. Each text was removed from the pool of texts after three subjects had done it. As a result of the experiment we had three human-selected REs for each of the MSR slots. There were a total of 764 MSR slots in this set of texts (an average of 8.4 per text).

There was a considerable amount of agreement among the subjects, despite the fact that there were on average 9.5 different REs to choose from for each MSR slot[5]. Table 1 shows an overview of the agreement figures. In just 8.9% of MSRs, all three subjects chose a different RE, whereas in 50.1% of MSRs, all subjects chose exactly the same RE. In 64.9% of cases the subjects all made the same decision about whether to pronominalise or not, and in 95.3% of cases they all agreed about whether to realise the MSR or not (this does, however, include a large number of cases where the non-realised reference is not grammatical, as e.g. in the example in Figure 1).

To assess agreement between the subjects and the corpus texts we computed the average of the three pairwise agreement figures, shown in Table 2. The average agreement figures here are somewhat higher than those in Table 1.

6 Automatically Chosing MSREs

We conducted experiments to obtain baseline results for automatically choosing among a given set of REs. The task definition was the same as in the

---

[5]Not all available choices are guaranteed to be grammatical, since they are not generated by a grammar-based component.

Total MSRs	764
Average MSRs per file	8.4
All three different	8.9%
All three exactly same	50.1%
All three same pronominalisation decision	64.9%
All three same non-realisation decision	95.3%

Table 1: (Dis)agreement among subjects in Choosing MSR Experiment.

Total MSRs	764
Average MSRs per file	8.4
Average pairwise agreement	65.0%
Same pronominalisation decision (avg)	76.7%
Same non-realisation decision (avg)	97.2%

Table 2: (Dis)agreement between subjects in Choosing MSR Experiment and corpus texts.

human experiments, i.e. we deleted all MSRs and replaced them with lists of possible REs.

Our goal was to determine the accuracy that can be obtained by making the most frequent choice in a given context. This kind of baseline has been shown to be very powerful for example in word sense disambiguation and POS-tagging. In POS-tagging, each word is tagged with its (generally) most frequent POS-tag; in WSD, each ambiguous word is tagged with its most frequent sense.

6.1 Automatic classification of REs

Methods for automatically choosing from a previously unseen set of REs need to map the REs to a generalized representation/classification that allows one to apply statistics obtained from the training corpus to new sets of REs. We devised a general classification scheme for REs which is based on the notion of default RE (see Section 4), RE length relative to length of default RE, and generally identifiable linguistic features (such as presence of a determiner and pronouns). The scheme distinguishes the following types and subtypes of REs:

1. Default name of the main subject of the article which we set to the title for each entry (e.g. *United Kingdom* for the entry on the United Kingdom of Britain and Northern Ireland).

2. Pronoun: (a) personal, (b) relative, (c) possessive.

3. REs with determiner: subcategorised according to length relative to default RE, length of the default +/- X, $1 \leq X \leq 6$.

4. Any other REs, subcategorised according to length relative to default RE: length of the default +/- X, $0 \leq X \leq 6$.

The idea in taking length into account is that this may enable us to capture length-related phenomena such as the fact that references to the same object in a discourse tend to become shorter over the course of a text (known as *attenuation* in the psycholinguistic literature).

6.2 Frequencies of MSR types in training set

We determined the frequencies of the above RE types in the training set as a whole, and individually for each subdomain (Rivers, Cities, Countries, People) and for each syntactic MSR type (Subjects, Objects, Possessives), as shown in Table 3. There are interesting differences in frequencies between different subdomains. Pronouns are overall the most frequent type of RE in our corpus, accounting for nearly half of all REs, but are more dominant in some subdomains than others: percentages range from 28% (Cities) to 63% (People). The default name (which we set to the entry title) is the second most frequent type overall, accounting for between 2% (Rivers[6]) and 37% (Cities).

REs that contain determiners are very rare in the People subdomain. REs shorter than the default are far more frequent in People (where reference by surname alone is common) than the other subdomains.

6.3 Most frequent choice selection

We tested most frequent choice selection on the test set (the same set of texts as was used in the human experiment) using four different ways of subdividing the corpus and calculating frequencies (S1–S4 below). For each corpus subdivision we ranked the RE types given above (Section 6.1) according to their frequency of occurrence in the corpus subdivision (these rank lists are referred to as *frequency lists* below). The four ways of subdivding the corpus were as follows:

S1. All texts, resulting in a single, global frequency list;

S2. Texts divided according to subdomain, resulting in four frequency lists (cities, countries, rivers, people);

S3. Texts divided according to MSR type, resulting in three frequency lists (subjects, objects, possessives);

S4. Texts divided according to both subdomain and MSR type, resulting in 12 frequency lists (one for each combination of subdomain and MSR type).

---

[6]The title in River entries often includes the word 'river', e.g. *Amazon River* whereas in REs in the texts it is rare.

length	Default	Pronoun (all)	RE +det $> d$	RE +det $< d$	RE +/-det $= d$	Other RE $> d$	Other RE $< d$
All texts (7277)	1491	3372	601	91	492	184	1046
All city texts (1735)	666	483	273	15	183	26	89
All country texts (1469)	521	506	227	57	112	6	40
All river texts (572)	13	245	98	10	143	3	50
All people texts (3501)	291	2138	3	9	54	149	867
All subject MSREs (4940)	1241	1863	398	50	364	171	853
All object MSREs (681)	184	148	129	31	102	13	74
All possessive MSREs (1656)	66	1361	74	10	26	0	119

Table 3: Training set frequencies of different RE types, computed for entire training set, subdomains and syntactic MSR types; d = length of default name.

	All	Cities	Countries	Rivers	People
All	29.6% (757)	49.7% (141)	36.7% (191)	4.2% (24)	57.1% (182)
Subject MSREs	34.8% (523)	49.1% (110)	43.0% (142)	29.4% (17)	42.1% (254)
Object MSREs	42.3% (52)	61.9% (16)	43.8% (16)	0% (2)	46.2% (13)
Possessive MSREs	85.2% (182)	50.0% (33)	90.9% (33)	80.0% (5)	86.6% (134)

Table 4: Test set results (in percent) obtained with several most frequent choice strategies 'trained' on different subsets of the training set.

This gave us 20 frequency lists in total which we applied to RE selection as follows. First, the alternative REs given in the test set inputs were classified with the scheme described in Section 6.1. Then the RE classified as belonging to the RE type at the top of the frequency list was selected. If no alternative was in the top RE category, we backed off to the second most frequent category, and so on.

Table 4 shows the percentages of correct decisions over the test set. The results clearly show that overall performance improves as more knowledge about the tasks is included. Subset sizes are shown in brackets in each cell, as they are informative: e.g. of the two objects MSRs in Rivers in the test set, neither was in the most frequent Object/River type according to the training set.

The 'global' accuracy figure (All/All) achieved with the frequency list computed from the entire training set is 29.6%; for the other sets, accuracy ranges from the very low 4.2% (All/River) to the very high 90.9% (Possessive/Country).

The more we know about what kind of MSR we are looking for, the better we can do. As computed on the entire test set, if we know nothing in addition to it being an RE, then we get 29.6%; if we (only) know whether the referent is a river, city, country or person, this figure rises to 48.9%; if we (only) know whether we are looking for a subject, object or possessive RE, then we get 47.4%. If we know both subdomain and MSR type, then we get as much

as 53.9%. This is still considerably less than the 65% achieved on average by the human subjects (Table 2), but it is a very strong baseline.

7 Further research

The distinct task we are planning to address in the immediate future is how well we can predict choice of REs using only input that is derivable from the full-text context, as would be required e.g. for text summarisation. The most frequent choice results presented in this paper represent baselines in this respect. In future work, we intend to look at more sophisticated approaches, including linguistic features (POS-tags, grammatical relations, etc.); optimal sequences of REs (e.g. modelled by n-gram models); and determining the current topic to decide which aspects of a referent are in focus (as described in Section 4).

We will also extend our annotation of the corpus texts in various ways, initially focussing on syntactic annotations such as POS-tags and dependencies. We also plan to look at annotating (or automatically identifying) potential distractors.

8 Conclusion

In this paper we presented and described our corpus of introductory encyclopaedic texts in which we have annotated three types of reference to the main subject. We described an experiment with human subjects in which we found that the subjects agreed

in their choice to a considerable degree. In our experiments with automatic RE selection we found that the simple strategy of selecting the most frequent type of RE provides a strong baseline, particularly if information regarding subdomain type and syntactic type of RE is included.

Acknowledgments

References

Bernd Bohnet and Robert Dale. 2005. Viewing referring expression generation as search. In *Proceedings of IJCAI'05*, pages 1004–1009.

Hua Cheng, Massimo Poesio, Renate Henschel, and Chris Mellish. 2001. Corpus-based np modifier generation. In *Proceedings of NAACL 2001*.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Robert Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*.

Robert Dale. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. Bradford Books, MIT Press, Cambridge, MA.

Surabhi Gupta and Amanda Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the 1st Workshop on Using Copora in Natural Language Generation*, pages 1–6, Brighton, UK.

Renate Henschel, Hua Cheng, and Massimo Poesio. 2000. Pronominalization revisited. In *Proceedings of COLING'00*, pages 306–312.

Rodney Huddleston and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Pamela W. Jordan and M. Walker. 2000. Learning attribute selections for non-pronominal expressions. In *Proceedings of ACL'00*.

Pamela W. Jordan. 2002. Contextual influences on attribute selection for repeated descriptions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. CSLI, Stanford, CA.

Emiel Krahmer and Mariet Theune. 2002. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 223–264. CSLI, Stanford, CA.

Kathy McCoy and Michael Strube. 1999. Generating anaphoric expressions: Pronoun or definite description. In *Proceedings of the ACL'99 Workshop on Reference and Discourse Structure*, pages 63–71.

Massimo Poesio. 2000. Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In *Proceedings of LREC 2000*.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proc. ACL'04 Discourse Annotation Workshop*.

Ehud Reiter and Robert Dale. 1992. A fast algorithm for the generation of referring expressions. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 232–238, Nantes, France, 23-28 August.

Advaith Siddharthan and Ann Copestake. 2004. Generating referring expressions in open domains. In *Proc. of ACL-04*.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Conference on Natural Language Generation*, pages 130–132, Sydney, Australia, July.

Kees van Deemter. 2002. Generating referring expressions: Boolean extensions of the Incremental Algorithm. *Computational Linguistics*, 28(1):37–52.

Jette Viethen and Robert Dale. 2006. Towards the evaluation of referring expression generation. In *Proc. of the 4th Australasian Language Technology Workshop (ALTW'06)*, pages 115–122.