Multilingual robust anaphora resolution

Ruslan Mitkov
School of Languages and European Studies
University of Wolverhampton
Stafford Street
Wolverhampton WV1 1SB, United Kingdom
Email R.Mitkov@wlv.ac.uk

Lamia Belguith
LARIS - FSEG
University of Sfax
B.P. 1088
3018 Sfax, Tunisia
Email belguith.lamia@planet.tn

Malgorzata Stys
Computer Laboratory
University of Cambridge
New Museums Site, Pembroke Street
Cambridge CB2 3QG
United Kingdom
Email Malgorzata.Stys@cl.cam.ac.uk

Abstract

Most traditional approaches to anaphora resolution rely heavily on linguistic and domain knowledge. One of the disadvantages of developing a knowledge-based system, however, is that it is a very labour-intensive and time-consuming task. This paper presents a robust, knowledge-poor approach to resolving pronouns in technical manuals. This approach is a modification of the practical approach (Mitkov 1998a) and operates on texts pre-processed by a part-of-speech tagger. Input is checked against agreement and a number of antecedent indicators. Candidates are assigned scores by each indicator and the candidate with the highest aggregate score is returned as the antecedent. We propose this approach as a platform for multilingual pronoun resolution. The robust approach was initially developed and tested for English, but we have also adapted and tested it for Polish and Arabic. For both languages, we found that adaptation required minimum modification and that further, even if used unmodified, the approach delivers acceptable success rates. Preliminary evaluation reports high success rates in the range of and over 90%

1. Introduction: robust, knowledge poor anaphora resolution and multilingual NLP

For the most part, anaphora resolution has focused on traditional linguistic methods (Carbonell & Brown 1988; Carter 1987; Hobbs 1978; Ingria & Stallard 1989; Lappin & McCord 1990; Lappin & Leass 1994; Mitkov 1994; Rich & LuperFoy 1988; Sidner 1979; Webber 1979). However, to represent and manipulate the various types of linguistic and domain knowledge involved requires considerable human input and computational expense.

While various alternatives have been proposed, making use of e.g. neural networks, a situation semantics framework, or the principles of reasoning with uncertainty (e.g. Connoly et al. 1994; Mitkov 1995; Tin & Akman 1995), there is still a strong need for the development of robust and effective strategies to meet the demands of practical NLP systems, and to enhance further the automatic processing of growing language resources.

Several proposals have already addressed the anaphora resolution problem by deliberately limiting the extent to which they rely on domain and/or linguistic knowledge (Baldwin 1997; Dagan & Itai 1990; Kennedy & Boguraev 1996; Mitkov 1998; Nasukawa 1994; Williams et al. 1996). Our work is a continuation of these latest trends in the search for inexpensive, rapid and reliable procedures for anaphora resolution. It shows how pronouns in a specific genre can be resolved quite successfully without any sophisticated linguistic knowledge or even without parsing, benefiting instead from corpus-based NLP techniques such as sentence splitting and part-of-speech tagging.

On the other hand, none of the projects reported so far, has looked at the multilingual aspects of the approaches that have been developed, or, in particular, how a specific approach could be used or adapted for other languages. Furthermore, in addition to the monolingual orientation of all approaches so far developed, most of the work has concentrated on pronoun resolution in one language alone (English).

While anaphora resolution projects have been reported for French (Popescu-Belis & Robba 1997, Rolbert 1989), German (Dunker & Umbach 1993; Fischer et al. 1996; Leass & Schwall 1991; Stuckardt 1996; Stuckardt 1997), Japanese (Mori et al. 1997; Nakaiwa & Ikehara 1992; Nakaiwa & Ikehara 1995; Nakaiwa et al. 1995; Nakaiwa et al. 1996; Wakao 1994), Portuguese (Abraços & Lopes 1994), Swedish (Fraurud, 1988) and Turkish (Tin & Akman, 1994), the research on languages other than English constitutes only a small part of all the work in this field.

In contrast to previous work in the field, our project has a truly multilingual character. We have developed a knowledge-poor, robust approach which we propose as a platform for multilingual pronoun resolution in technical manuals. The approach was initially developed and tested for English, but we have also adapted and tested it for Polish and Arabic. We found that the approach could be adapted with minimum modifications for both languages and further, even if used without any modification, it delivers acceptable success rates. Evaluation shows a success rate of 89.7% for English, 93.3% for Polish and 95.2% for Arabic.[1]

2. The approach: general overview

With a view to avoiding complex syntactic, semantic and discourse analysis, we developed a robust, knowledge-poor approach to pronoun resolution which does not make use of parsing, syntactic and semantic constraints or any other form of linguistic or non-linguistic knowledge. Instead, we rely on the efficiency of sentence segmentation, part-of-speech tagging, noun phrase identification and the high performance of the antecedent indicators (knowledge is limited to a small noun phrase grammar, a list of terms, a list of (indicating) verbs, a list of genre-specific synonyms, and a set of antecedent indicators).

The core of the approach lies in activating a list of multilingual[2] "antecedent indicators" after filtering candidates (from the current and two preceding sentences) on the basis of gender and number agreement. Before that, the text is pre-processed by a sentence splitter which determines the sentence boundaries, a part-of-speech tagger which identifies the parts of the speech and a simple phrasal grammar which detects the noun phrases (In addition, in the case of complex sentences, heuristic "clause identification" rules track the clause boundaries). Non-anaphoric occurrences of "it" in constructions such as "It is important", "It is necessary" etc., are eliminated by a "referential filter".

After passing the "agreement filter", the genre-specific antecedent indicators are applied to the remaining candidates (see section 2.2). The noun phrase with the highest aggregate score is proposed as antecedent; in the rare event of a tie, priority is given to the candidate with the higher score for immediate reference. If immediate reference has not been identified, then priority is given to the candidate with the best collocation pattern score. If this does not help, the candidate with the higher score for indicating verbs is preferred. If still no choice is possible, the most recent from the remaining candidates is selected as the antecedent.

2.1 Agreement filter

The detected noun phrases (from the sentence where the anaphor is situated and the two preceding sentences, if available) are passed on to a gender and number agreement test. In English, however, there are certain collective nouns which do not agree in number with their antecedents (e.g. "government", "team", "parliament" etc. can be referred to by "they"; equally some plural nouns such as "data" can be referred to by "it") and are exempted from the agreement test. For this purpose we have drawn up a comprehensive list of all such cases; to our knowledge, no other computational treatment of pronominal anaphora resolution has addressed the problem of "agreement exceptions".

The gender and number agreement of an anaphor and its antecedent in Polish is compulsory. Polish gender distinctions are much more diverse than in English (e.g. feminine and masculine do not apply to a restricted number of nouns). Moreover, one pronominal form can potentially refer to nouns of different gender. For instance, the singular genitive form "jego" can equally well refer to either masculine or neuter nouns. In addition, certain pronouns such as the accusative form "je" can refer to either singular neuter or plural feminine nouns. Finally, unlike English, zero anaphors (in subject position) are typical in Polish in declarative sentences.

Agreement rules in Arabic are different. For instance, a set of non-human items (animals, plants, objects) is referred to by a singular feminine pronoun. Since Arabic is an agglutinative language, the pronouns may appear as suffixes of verbs, nouns (e.g. in the case of possessive pronouns) and prepositions. In particular, in the genre of technical manuals there are five "agglutinative" pronouns. The pronoun "ho" is used to refer to singular masculine persons and

[1] Given that the evaluation of the English version was more extensive, the figures for English are expected to be statistically more representative.

[2] We term the antecedent indicators "multilingual" because they work well not only for English, but also for other languages (in this case Arabic and Polish).

objects, while "ha" refers to singular feminine ones. There are three plural anaphoric pronouns: "homa" which refers to a dual number (a set of two elements) of both masculine and feminine nouns, "hom" which refers to a plural number (a set of more than two elements) of masculine nouns and "honna" which refers to a plural number of feminine

2.2 Antecedent indicators

Antecedent indicators (preferences) play a decisive role in tracking down the antecedent from a set of possible candidates. Candidates could be given preferential treatment, or not, from the point of view of each indicator and assigned a score (-1, 0, 1 or 2) accordingly; the candidate with the highest aggregate score is proposed as the antecedent. The antecedent indicators have been identified on the basis of empirical studies of numerous hand-annotated technical manuals (referential links had been marked by human experts). These indicators can be related to salience (definiteness, givenness, indicating verbs, indicating noun phrases, lexical reiteration, section heading preference, "non-prepositional" noun phrases, relative pronoun), to structural matches (collocation, immediate reference, sequential instructions), to referential distance or to preference of terms. Whilst some of the indicators are more genre-specific (term preference) and others are less genre-specific ("immediate reference", "sequential instructions" and to a much lesser extent "indicating noun phrases"), the majority of them appear to be genre-independent. In the following we shall outline the indicators used and shall illustrate some of them by examples (the indicators are used in the same way for English, Polish and Arabic unless otherwise specified).

Definiteness

Definite noun phrases in previous sentences are more likely antecedents of pronominal anaphors than indefinite ones (definite noun phrases score 0 and indefinite ones are penalised by -1). In English we regard a noun phrase as definite if the head noun is modified by a definite article, or by demonstrative or possessive pronouns. This rule is ignored if there are no definite articles, possessive or demonstrative pronouns in the paragraph (this exception is taken into account because some English user's guides tend to omit articles).

Since in Polish there are no definite articles, definiteness is signalled by word order, demonstrative pronouns or repetition.

In Arabic, definiteness occurs in a richer variety of forms (Galaini 1992). In addition to the definiteness triggered by the definite article "al" (the), demonstrative and possessive pronouns, a noun phrase in Arabic is also regarded as definite if it is followed by a definite noun/noun phrase[3]. For example, the noun phrase "kitabu al-rajuli" (lit. book the man) which means "the book of the man", is considered definite since the non-definite noun "kitabu" (book) is followed by the definite noun "al-rajoli" (the man). This form of definiteness is called in Arabic "Al-ta'rif bi-al-idhafa" (definiteness by addition).

Givenness

Noun phrases in previous sentences representing the "given information" (theme)[4] are deemed good candidates for antecedents and score 1 (candidates not representing the theme score 0). In a coherent text (Firbas 1992), the given or known information, or theme, usually appears first, and thus forms a co-referential link with the preceding text. The new information, or rheme, provides some information

Indicating verbs

If a verb is a member of the Verb_set = {discuss, present, illustrate, identify, summarise, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyse, synthesise, study, survey, deal, cover}, we consider the first NP following it as the preferred antecedent (scores 1 and 0). Empirical evidence suggests that because of the salience of the noun phrases which follow them, the verbs listed above are particularly good indicators.

The Verb_set in Polish contains the Polish equivalents of the above verbs and their synonyms.

Indicating noun phrases

If the head of the NP preceding the verb is the noun "chapter", "section", "table" then consider the NP following the verb as the preferred antecedent (scores 1 and 0)

The last two preferences can be illustrated by the example:

> This table shows a minimal configuration$_i$; it$_i$ does not leave much room for additional applications or other software for which you may require additional swap space.

[3]There are other forms of definiteness in Arabic which we shall not discuss in this paper since they are not typical of technical manuals.

[4]We use the simple heuristics that the given information is the first noun phrase in a non-imperative sentence.

Lexical reiteration

Lexically reiterated items are likely candidates for antecedent (a NP scores 2 if is repeated within the same paragraph twice or more, 1 if repeated once and 0 if not). Lexically reiterated items include repeated synonymous noun phrases which may often be preceded by definite articles or demonstratives. Also, a sequence of noun phrases with the same head counts as lexical reiteration (e.g. "toner bottle", "bottle of toner", "the bottle").

Section heading preference

If a noun phrase occurs in the heading of the section, part of which is the current sentence, then we consider it as the preferred candidate (1, 0).

"Non-prepositional" noun phrases

A "pure", "non-prepositional" noun phrase is given a higher preference than a noun phrase which is part of a prepositional phrase (0, -1)

> Insert the cassette$_i$ into the VCR making sure it$_i$ is suitable for the length of recording.

Here "the VCR" is penalised (-1) for being part of the prepositional phrase "into the VCR".

This preference can be explained in terms of salience from the point of view of the centering theory. The latter proposes the ranking "subject, direct object, indirect object" (Brennan et al. 1987) and noun phrases which are parts of prepositional phrases are usually indirect objects.

This criterion was extended in Polish to frequently occurring genitive constructions (e.g. liczba komputerow = number of computers). Nouns which are part of such genitive constructions and which are not in genitive form are penalised by "-1".

In Arabic the antecedent and the anaphor can belong to the same prepositional phrase (see next section). Therefore, we have modified this indicator for the "Arabic version" accordingly: if an NP belongs to a prepositional phrase which doesn't contain the anaphor, we penalise it by -1; otherwise we do not assign any score to it (0).

Relative pronoun indicator

This indicator is used only in the Arabic version and is based on the fact that the first anaphor following a relative pronoun refers exclusively to the most recent NP preceding it which is considered as the most likely antecedent (2,0).

Example:

> Al-tahakkok min tahyiat al-moakkit
> Yomkino-ka a'rdh tahyiat moakkitoka li-at-tahakkok mina al-baramij$_i$ al-lati targhabo fi tasjili-ha$_i$.
> (Literal translation)
> Checking the Timer Settings
> You can display your timer settings to confirm the programmes$_i$ that you wish to recording it$_i$.
> Checking the Timer Settings
> You can display your timer settings to confirm the programmes you wish to record.

In this example the pronoun "ha" (it) is the first pronominal anaphor which follows the relative pronoun "al-lati" (that) and refers to the non-animate feminine plural "al-baramij" (the programmes; for agreement rules in Arabic see section 2.1) which is the most recent NP preceding "al-lati".

Collocation pattern preference

This preference is given to candidates which have an identical collocation pattern with a pronoun (2,0). The collocation preference here is restricted to the pattern "noun/pronoun, verb" or "verb, noun/pronoun" (owing to lack of syntactic information, this preference is somewhat weaker than the collocation preference described in (Dagan & Itai 1990).

> Press the key$_i$ down and turn the volume up... Press it$_i$ again.

The collocation pattern preference in Arabic has been extended to patterns "(un)V-NP/anaphor", i.e. verbs with a "undoing action" meaning are considered for the purpose of our approach to fall into collocation patterns along with their "doing action" counterparts. This extended new rule would help in cases such as "Loading a cassette or unloading it". This rule is soon to be integrated into the English and Polish versions.

Immediate reference

In technical manuals the "immediate reference" clue can often be useful in identifying the antecedent. The heuristics used is that in constructions of the form "...(You) V_1 NP ... con (you) V_2 it (con (you) V_3 it)", where con $\in$ {and/or/before/after...}, the noun phrase immediately after V_1 is a very likely candidate for antecedent of the pronoun "it" immediately following V_2 and is therefore given preference (scores 2 and 0).

This preference can be viewed as a modification of the collocation preference. It is also quite frequent with imperative constructions.

> To print the paper, you can stand the printer$_i$ up or lay it$_i$ flat.

To turn on the printer, press the Power button$_i$ and hold it$_i$ down for a moment.
Unwrap the paper$_j$, form it$_j$ and align it$_j$, then load it$_j$ into the drawer.

Sequential instructions

This new antecedent indicator has recently been incorporated for Arabic but it works equally well for English and is to be implemented in the English version soon as well. It states than in sequential instructions of the form "To V_1 NP_1, V_2 NP_2. (Sentence). To V_3 it, V_4 NP_4" the noun phrase NP_1 is the likely antecedent of the anaphor "it" (NP_1 is assigned a score of 2).

Example:

To turn on the video recorder, press the red button. To programme it, press the "Programme" key.
To turn the TV set ON, press the mains ON/OFF switch. The power indicator illuminates to show that the power is on. To turn the TV set off, press it again.

Referential distance

In English complex sentences, noun phrases in the previous clause[5] are the best candidate for the antecedent of an anaphor in the subsequent clause, followed by noun phrases in the previous sentence, then by nouns situated 2 sentences further back and finally nouns 3 sentences further back (2, 1, 0, -1). For anaphors in simple sentences, noun phrases in the previous sentence are the best candidate for antecedent, followed by noun phrases situated 2 sentences further back and finally nouns 3 sentences further back (1, 0, -1)

Since we found out that in Arabic the anaphor is more likely to refer to the most recent NP, the scoring system for Arabic gives a bonus to such candidates: the most recent NP is assigned a score of 2, the one that precedes it immediately 1 and the rest 0.

Term preference

NPs representing terms in the field are more likely to be the antecedent than NPs which are not terms (score 1 if the NP is a term and 0 if not).

As already mentioned, each of the antecedent indicators assigns a score with a value $\in \{-1, 0, 1, 2\}$. These scores have been determined experimentally on an empirical basis and are constantly being updated. Top symptoms like "lexical reiteration" assign score "2" whereas "non-prepositional" noun phrases are given a negative score of "-1". We should point out that the antecedent indicators are preferences and not absolute factors. There might be cases where one or more of the antecedent indicators do not "point" to the correct antecedent. For instance, in the sentence "Insert the cassette into the VCR$_i$ making sure it$_i$ is turned on", the indicator "non-prepositional noun phrases" would penalise the correct antecedent. When all preferences (antecedent indicators) are taken into account, however, the right antecedent is still very likely to be tracked down - in the above example, the "non-prepositional noun phrases" heuristics (penalty) would be overturned by the "collocational preference" heuristics.

The antecedent indicators have proved to be reasonably efficient in assigning the right antecedent and our results show that for the genre of technical manuals they may be no less accurate than syntax- and centering-based methods (see Mitkov 1998b). The approach described is not dependent on any theories or assumptions; in particular, it does not operate on the assumption that the subject of the previous utterance is the highest-ranking candidate for the backward-looking center - an approach which can sometimes lead to incorrect results. For instance, most centering-orientated methods would propose "the utility" incorrectly as the antecedent of "it" in the sentence "The utility (CDVU) shows you a LIST4250, LIST38PP, or LIST3820 file on your terminal for a format similar to that in which it will be printed" because of the preferential treatment of the subject as the most salient candidate (e.g. RAP, see Dagan et al. 1995). The "indicating verbs" preference of our approach, however, would give preference to the correct antecedent "LIST4250, LIST38PP, or LIST3820 file".

3. Evaluation

For practical reasons, the approach presented does not incorporate syntactic and semantic knowledge (other than a list of domain terms) and it is not realistic to expect its performance to be as good as an approach which makes use of syntactic and constraints and preferences. The lack of syntactic information, for instance, means giving up c-command constraints and subject preference (or on other occasions object preference, see Mitkov 1995) which could be used in center tracking. Syntactic parallelism, useful in discriminating between identical pronouns on the basis of their syntactic function, also has to be forgone. Lack of semantic knowledge rules out the use of verb semantics and semantic parallelism. Our evaluation, however, suggests that much less is lost than might be feared. In fact, our evaluation shows that the results are comparable to and

[5]Identification of clauses in complex sentences is done heuristically.

even better than syntax-based methods (Lappin & Leass 1994). The evaluation results also show superiority over other knowledge-poor methods (Baldwin 1997; see also below)[6]. We believe that the good success rate is due to the fact that a number of antecedent indicators are taken into account and no factor is given absolute preference. In particular, this strategy can often override incorrect decisions linked with strong centering preference (see 2.2) or syntactic and semantic parallelism preferences (Mitkov 1998b).

We have carried out evaluations on sample texts from technical user's guides both for English and Arabic and the results show comparable success rates. The success rate for Arabic is slightly higher and we should mention that in addition to tuning the approach for Arabic, the "Arabic improved" version uses 2 new indicators recently introduced which have not been included in the "Robust English" version yet.

3.1 English

The first evaluation exercise for English (Mitkov & Stys 1997) was based on a random sample text from a technical manual (Minolta 1994). There were 71 pronouns in the 140 page technical manual; 7 of the pronouns were non-anaphoric and 16 exophoric. The resolution of anaphors was carried out with a success rate of 95.8%. The approach being robust (an attempt is made to resolve each anaphor and a proposed antecedent is returned), this figure represents both "precision" and "recall" if we use the MUC terminology. To avoid any terminological confusion, we shall therefore use the more neutral term "success rate" while discussing the evaluation.

We conducted a second evaluation[7] of the robust approach on a different set of English sample texts from the genre of technical manuals (47-page Portable Style-Writer User's Guide (Stylewriter 1994). Out of 223 pronouns in the text, 167 were non-anaphoric (deictic and non-anaphoric "it"). The evaluation carried out was manual to ensure that no added error was generated (e.g. due to possible wrong sentence/clause detection or POS tagging). Another reason for doing it by hand is to ensure a fair comparison with other knowledge-poor methods (Baldwin 1997), which not being available to us, had to be hand-simulated.

The second evaluation indicated an 83.6% success rate for our robust approach. Baldwin's CogNIAC scored 75% on the same data, while J. Hobb's algorithm achieved 71% (Mitkov 1998b).

On the basis of both evaluation experiments a success rate of 89.7% could be regarded as a statistically more representative figure for the performance of "English version" of the robust approach[8]. In addition, our evaluation results indicate 82% "critical success rate", which we consider quite a satisfactory score (for definition of the concept "critical success rate" which is limited to the evaluation of the so-called "critical cases" - the resolution of "tough" anaphors which have already passed the agreement filter, see Mitkov 1998b). Finally, in order to evaluate the effectiveness of the approach and to explore whether or by how much it is superior to the baseline models for anaphora resolution, we also tested the sample texts on (i) a Baseline Model which checks agreement in number and gender and, where more than one candidate remains, picks as antecedent the most recent subject matching the gender and number of the anaphor and (ii) a Baseline Model which picks as antecedent the most recent noun phrase that matches the gender and number of the anaphor. The evaluation results suggest a success rate of 48.55% for the first baseline model and a success rate 65.95% for the second (Mitkov 1998b).

If we regard as "discriminative power" of each antecedent indicator the ratio "number of successful antecedent identifications when this indicator was applied"/"number of applications of this indicator" (for the non-prepositional noun phrase and definiteness being penalising indicators, this figure is calculated as the ratio "number of unsuccessful antecedent identifications"/"number of applications"), the immediate reference emerges as the most discriminative indicator (100%), followed by non-prepositional noun phrase (92.2%), collocation (90.9%), section heading (61.9%), lexical reiteration (58.5%), givenness (49.3%), term preference (35.7%) and referential distance (34.4%). The relatively low figures for the majority of indicators should not be regarded as a surprise: firstly, we should bear in mind that in most cases a candidate was picked (or rejected) as an antecedent on the basis of applying a number of different indicators and secondly, that most anaphors had a relatively high number of candidates for antecedent.

In terms of frequency of use ("number of non-zero applications"/"number of anaphors"), the most frequently used indicator proved to be referential distance used in 98.9% of the cases, followed by term preference (97.8%), givenness (83.3%), lexical reit-

[6] This applies to the genre of technical manuals; for other genres results may be different

[7] We are indebted to Lowenna Ansell for carrying out the second evaluation

[8] Please note that we have recently modified some of the rules/added some more rules but we have not evaluated the improved English version yet.

eration (64.4%), definiteness (40%), section heading (37.8%), immediate reference (31.1%) and collocation (11.1%). As expected, the most frequent indicators were not the most discriminative ones.

3.2 Arabic

We evaluated the robust approach for Arabic operating in two modes: the first mode consisted of using the robust approach directly, without any adaptation/modification for Arabic, whereas the second mode used an adapted/enhanced version which included modified rules (see section 2.2) designed to capture some of the specific aspects of Arabic plus a few new indicators.

The evaluation was based on 63 examples from a technical manual (Sony 1992). The first mode (i.e. using the robust approach without any adaptation for Arabic - this version is referred to as "Arabic direct" in the table below) reported a success rate of 90.5% (57 out of 63 anaphors were correctly resolved). Typical failures were examples in which the antecedent and the anaphor belonged to the same prepositional phrase:

> Tathhar al-surah fi awal kanat$_i$ ta-stakbilo-ha$_i$ fi mintakati-ka.
> Appears the-picture on first channel$_i$ you-receive-it$_i$ in area-your. (Literal translation)
> The picture appears when the first channel received in your area is detected.

Such failure cases were not detected in the improved version for Arabic in which the "non-prepositional phrase" rule was changed (see section 2.2).

Another typical problem which was rectified by changing the referential distance in Arabic was the case in which the anaphor appeared as part of a PP modifying the antecedent-NP:

> Kom bi-taghtiat thokb al-lisan bi-sharit plastic aw ista'mil kasit akhar$_i$ bi-hi$_i$ lisan al-aman.
> Cover slot the-tab with-tape plastic or use cassette another$_i$ in it$_i$ tab the- safety.
> Cover the safety tab slot with plastic tape, or use another cassette with a safety tab.

The candidates for antecedent in this example are the noun phrases "safety tab slot", "plastic tape" and "another cassette". If we use the robust approach without any modification, each candidate gets 2 for referential distance; the aggregate score for "safety tab slot" is 3, for "plastic tape" it is 2 and for "another cassette" is 2 as well (they all get an additional 1 score for "term preference"). Using the new referential distance scores, however, the correct candidate "another cassette" scores an aggregate of 2 as opposed to the other two candidates which are assigned an aggregate score of 1.

The second evaluation mode (evaluating the version adapted and improved for Arabic which is referred to as "Arabic improved" in the table below) reported a success rate of 95.2% (60 out of 63 anaphors were correctly resolved).

The evaluation for Arabic also showed a very high "critical success rate" as well. The robust approach used without any modification scored a "critical success rate" of 78.6%, whereas the improved Arabic version scored 89.3%.

The most discriminative indicators for Arabic proved to be immediate reference, collocation and sequential instructions with 100% discriminative power, followed by non-prepositional noun phrase (89.2%), term preference (82.1%), definiteness (78.6%), referential distance_score_2 (67.9%) and section heading (63.6%). The higher contribution of referential distance for Arabic is in tune with our empirical finding that referential distance is a more important indicator for Arabic than for English and that in particular, the most recent NPs in Arabic are more likely to be antecedents than in English (see section 2.2, indicator "referential distance").

The most frequently used indicators for Arabic were referential distance (100%, of which 34.6% with score 2 and 34.6% with score 1) and term preference (87.7%). Again, the most discriminative indicators could not be frequently used: collocation was applied in only 2.5% of the cases, whereas immediate reference and sequential instructions could be activated in 1.2% of the cases only.

3.3 Polish

The evaluation for Polish was based technical manuals available on the Internet (Internet Manual, 1994; Java Manual 1998). The sample texts contained 180 pronouns among which were 120 instances of exophoric reference (most being zero pronouns). The robust approach adapted for Polish demonstrated a high success rate of 93.3% in resolving anaphors.

Similarly to the evaluation for English, we compared the approach for Polish with (i) a Baseline Model which discounts candidates on the basis of agreement in number and gender and, if there were still competing candidates, selects as the antecedent the most recent subject matching the anaphor in gender and number (ii) a Baseline Model which checks agreement in number and gender and, if there were still more than one candidate left, picks up as the antecedent the most recent noun phrase that agrees with the anaphor.

The Polish version of our robust approach showed clear superiority over both Polish baseline models.

The first Baseline Model (Baseline Subject) was successful in only 23.7% of the cases, whereas the second (Baseline Most Recent) had a success rate of 68.4%. These results demonstrate the dramatic increase in precision, which is due to the use of antecedent tracking indicators.

The Polish version also showed a very high "critical success rate" of 86.2%. Used without any modification ("Polish direct"), the approach scored a 90% success rate.

The most discriminative antecedent indicators for Polish appear to be the sequential instructions, immediate reference and indicating verbs (100%), followed by referential distance (84.1%) and givenness (80 %).

The most frequently used indicators for Polish were definiteness (97.2% of the cases), referential distance (94.4%), givenness (61.1%) and non-prepositional noun phrase (52.8%). The least frequently used indicators proved to be indicating verbs (16.7%), lexical reiteration (13.9%) and immediate reference (2.8%).

The success rates obtained can be summarised as follows:

	Success rate
Robust English	89.7%
Polish direct	90%
Polish improved	93.3%
Arabic direct	90.5%
Arabic improved	95.2%

Table 1: Success rates of the robust approach

	Success rate
Baseline subject English	31.6% / 48.6%
Baseline most recent English	65.9%
Baseline subject Polish	23.7%
Baseline most recent Polish	68.4%

Table 2: Success rates of the baseline models

Since the approach is robust, the success rates equal both recall and precision except for "Baseline subject English": since there are cases in which "Baseline subject" may not be able to pick up an antecedent (e.g. paragraphs with zero subjects), this version can be measured in terms of both precision (the higher figure in table 2) and recall (the lower figure).

4. Future work

Future work includes adapting the approach for French, Spanish and Bulgarian as well as testing it on (and if necessary, modifying it to cover) a wider variety of genres. In addition, we plan to use the statistically-based multicriteria approach (Pomerol & Barbara-Romero, 1992) to fine-tune scoring.

5. Conclusion

We have described a genre-specific modification of the practical approach to pronoun resolution (Mitkov 1998a) and have shown its multilingual nature: we have adapted and tested the approach for Polish and Arabic. The evaluation reports success rates which are comparable to (and even better than) syntax-based methods and show superiority over other methods with limited knowledge.

References

Abraços, Jose & José G. Lopes. 1994. "Extending DRT with a focusing mechanism for pronominal anaphora and ellipsis resolution". *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, 1128-1132, Kyoto, Japan.

Baldwin, Breck. 1997. "CogNIAC: high precision coreference with limited knowledge and linguistic resources". *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, 38-45, Madrid, Spain.

Brennan, S., M. Fridman and C. Pollard. 1987. A centering approach to pronouns. *Proceedings of the 25th Annual Meeting of the ACL (ACL'87)*, 155-162. Stanford, CA, USA.

Carbonell, James G. & Ralf D. Brown. 1988. "Anaphora resolution: a multi-strategy approach". *Proceedings of the 12. International Conference on Computational Linguistics (COLING'88)*, Vol.I, 96-101, Budapest, Hungary.

Carter, David M. 1987. *Interpreting anaphora in natural language texts*. Chichester: Ellis Horwood

Connoly, Dennis, John D. Burger & David S. Day. 1994. "A Machine learning approach to anaphoric reference". *Proceedings of the International Conference "New Methods in Language Processing"*, 255-261, Manchester, United Kingdom.

Dagan, Ido & Alon Itai. 1990. "Automatic processing of large corpora for the resolution of anaphora references". *Proceedings of the 13theInternational Conference on Computational Linguistics (COLING'90)*, Vol. III, 1-3, Helsinki, Finland.

Dagan, Ido, John Justeson, Shalom Lappin, Herbert Leass & Amnon Ribak. 1995. Syntax and lexical statistics in anaphora resolution. *Applied Artificial Intelligence*, 9.

Dunker, Guido & Carla Umbach. 1993. *Verfahren zur Anaphernresolution in KIT-FAST*. Internal Report KIT-28, Technical University of Berlin.

Firbas, Jan. 1992. *Functional sentence perspective in written and spoken communication*. Cambridge: Cambridge University Press.

Fischer, Ingrid, Bernd Geistert & Günter Görz 1996. "Incremental anaphora resolution in a chart-based semantics construction framework using l-DRT". *Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution. Lancaster (DAARC)*, 235-244, Lancaster, UK.

Fraurud, Kari. 1988. "Pronoun Resolution in unrestricted text". *Nordic Journal of Linguistics* 11, 47-68

Galaini, Chikh Mustafa. 1992. *Jami'u al-durus al-arabiah* (Arabic lessons collection). Beirut: Manshurat al-maktabah al-asriyah (Modern library).

Hasan, Abbes. 1975. *Al-nahw al-wafi ma'a rabtihi bi-al-asalib al-rafiah wa al-hayah al-loghawiah al-mutajadidah* (Complete grammar referring to good styles and the changing language). Egypt: Dar al-ma'arif (Knowledge bookstore).

Hobbs, Jerry R. 1978 "Resolving pronoun references". *Lingua*, 44, 339-352.

Ingria, Robert J.P. & David Stallard. 1989. "A computational mechanism for pronominal reference". *Proceedings of the 27th Annual Meeting of the ACL*, 262-271, Vancouver, British Columbia.

Internet Manual. 1994. *Translation of Internet Manual Internet i okolice: Przewodnik po swiatowych sieciach komputerowych*. Tracy LaQuey, Jeanne C. Ryer Translated by Monika Zielinska, BIZNET Poland.

Java Manual. 1998. *Jezyk Java*. Clico, Krakow.

Kennedy, Christopher & Branimir Boguraev, 1996. "Anaphora for everyone: pronominal anaphora resolution without a parser". *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, 113-118. Copenhagen, Denmark.

Lappin, Shalom & Michael McCord. 1990. "Anaphora resolution in slot grammar". *Computational Linguistics*, 16:4, 197-212.

Lappin, Shalom & Herbert Leass. 1994. "An algorithm for pronominal anaphora resolution". *Computational Linguistics*, 20(4), 535-561.

Leass Herbert & Ulrike Schwall. 1991. *An anaphora resolution procedure for machine translation*. IBM Germany Science Center. Institute for Knowledge Based Systems, Report 172.

Minolta. 1994. *Minolta Operator's Manual for Photocopier EP5325*. Technical Manual Minolta Camera Co., Ltd., Business Equipment Division 3-13, 2-Chome, Azuchi, -Machi, Chuo-Ku, Osaka 541, Japan.

Mitkov, Ruslan. 1994. "An integrated model for anaphora resolution". *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, 1170-1176, Kyoto, Japan.

Mitkov, Ruslan. 1995. "Un uncertainty reasoning approach for anaphora resolution". *Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'95)*, 149-154, Seoul, Korea.

Mitkov, Ruslan. 1998a. "Pronoun resolution: the practical alternative". In T. McEnery, S. Botley(Eds) *Discourse Anaphora and Anaphor Resolution*. John Benjamins.

Mitkov, Ruslan. 1998b. "Evaluating anaphora resolution approaches" (forthcoming).

Mitkov, Ruslan & Malgorzata Stys. 1997. "Robust reference resolution with limited knowledge: high precision genre-specific approach for English and Polish". *Proceedings of the International Conference "Recent Advances in Natural Language Proceeding" (RANLP'97)*, 74-81. Tzigov Chark, Bulgaria.

Mori, Tatsunori, Mamoru Matsuo, Hiroshi Nakagawa. 1997. Constraints and defaults of zero pronouns in Japanese instruction manuals. *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, 7-13. Madrid, Spain.

Nakaiwa, Hiromi & Satoru, Ikehara. 1992. " Zero pronoun resolution in a Japanese-to-English Machine Translation system by using verbal semantic attributes". *Proceedings of 3rd Conference on Applied Natural Language Processing (ANLP'92)*, Trento, Italy.

Nakaiwa, Hiromi & Satoru, Ikehara. 1995. "Intrasentential resolution of Japanese zero pronouns in a Machine Translation system using semantic and pragmatic constraints". *Proceeedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'95)*, 96-105, Leuven, Belgium.

Nakaiwa, Hiromi, S. Shirai, Satoru Ikehara & T. Kawaoka. 1995. "Extrasentential resolution of Japanese zero pronouns using semantic and pragmatic constraints". *Proceedings of the AAAI 1995 Spring Symposium Series: Empirical methods in discourse interpretation and generation.*

Nakaiwa, Hiromi & Francis Bond, Takahiro Uekado & Yayoi Nozawa. 1996. "Resolving zero pronouns in texts using textual structure". *Proceedings of the International Conference "New Methods in Language Processing" (NeMLaP-2)*, Ankara, Turkey.

Nasukawa, Tetsuya. 1994. "Robust method of pronoun resolution using full-text information". *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, 1157-1163, Kyoto, Japan.

Pomerol, Jean-Charles & Sergio Barbara-Romero. 1992. *Choix multicritère dans l'entreprise: principes et pratique*. Paris: HERMES.

Popescu-Belis, Andrei & Isabelle Robba. 1997. "Cooperation between pronoun and reference resolution for unrestricted texts". *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, 30-37. Madrid, Spain.

Rich, Elaine & Susann LuperFoy. 1988. "An architecture for anaphora resolution". *Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-2)*, 18-24, Texas, U.S.A.

Rolbert, Monique. 1989. *Résolution de formes pronominales dans l'interface d'interogation d'une base de données*. Thèse de doctorat. Faculté des sciences de Luminy.

Sidner, Candy L. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. Technical Report No. 537. M.I.T., Artificial Intelligence Laboratory.

Sony. 1992. *Video cassette recorder*. Operating Instructions. Sony Corporation.

Stuckardt, Roland. 1996. "An interdependency-sensitive approach to anaphor resolution". *Proceedings of the International Colloquium on Discourse Anaphora and*

Anaphora Resolution. Lancaster (DAARC), 400-413. Lancaster, UK.

Stuckardt, Roland. 1997. "Resolving anaphoric references on deficient syntactic descriptions". *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, 30-37. Madrid, Spain.

Stylewriter 1994. *Portable StyleWriter*. User's guide. Apple Computers.

Tin, Erkan & Varol, Akman. 1994. "Situated processing of pronominal anaphora". *Proceedings of the KONVENS"94 Conference*, 369-378, Vienna, Austria.

Wakao, Takahiro. 1994. "Reference resolution using semantic patterns in Japanese newspaper articles". *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, 1133-1137. Kyoto, Japan.

Webber, Bonnie L. 1979. *A formal approach to discourse anaphora*. London: Garland Publishing.

Williams, Sandra, Mark Harvey & Keith Preston. 1996. "Rule-based reference resolution for unrestricted text using part-of-speech tagging and noun phrase parsing". *Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution (DAARC)*, 441-456. Lancaster, UK.