Mutual information applied to coreference: an information-theoretic evaluation measure

Andrei Popescu-Belis
ISSCO, University of Geneva
Switzerland

Unpublished manuscript, November 2002.

Abstract

The accurate communication of references to entities is viewed here from an information-theoretic perspective. The successful transmission of references amounts to an similar distribution of coreferent expressions for the sender and for the receiver of a discourse, which is an alternative view with regard to equivalence classes for the coreference relation. Mutual information can be used to define entropy-based recall and precision for the evaluation of reference resolution. This measure and others proposed in this domain are illustrated here on various texts to determine their relevance.

Introduction

Reference resolution has been a constantly active research topic in the past decades. Despite several evaluation campaigns (*e.g.*, the Message Understanding Conferences), and many published papers, there seems to be no general agreement on a standard evaluation measure. This article proposes a measure based on a sound theoretical base, namely information theory.

After an overview of the problem, we will introduce our information-theoretic model and compare it with others. We then define our measure based on *mutual referring information* between the speaker and the hearer, and establish some theoretical properties. Finally, we compare our proposal with others, by applying it to a range of real or toy examples.

1 Aspects of reference in discourse

Specific fragments of utterances in discourse are devoted to a particular function: they evoke *entities*, functioning thus as *referring expressions* (henceforth, *REs*). We call *discourse entities* (henceforth, *DEs*) the conceptual structures that REs in a discourse refer to, following, *e.g.*, (Cristea, Postolache, Dima and Barbu 2002, Grosz, Joshi and Weinstein 1995). Discourse entities correspond generally to physical or mental objects, such as persons, things, ideas, but also to "reified" events, relations or properties.

The relation between REs and DEs is best named *specification*, following Sidner (1983) among others (*reference* denotes the relation to external-world entities). Specification occurs in the minds of both the speaker and the hearer of a discourse, but in opposite directions: *DE* → *RE* for the speaker, *vs. RE* → *DE* for the hearer.

Two important linguistic concepts related to reference are *coreference* and *anaphora*. Coreference is the relation that holds between two REs that specify the same DE (also called co-specification). Anaphora is a relation between an *antecedent* RE and an *anaphoric* RE, and holds, in its broadest sense, when the anaphor cannot be fully interpreted from the point of view of reference without making use of the antecedent. Anaphora can occur without coreference (for instance in the case of bridging or associative anaphora) and conversely, REs can be coreferent but not anaphoric (*e.g.*, consecutive uses of the same name)—see also (van Deemter and Kibble 2000, Vieira and Poesio 2000). Pronouns are often anaphoric: by virtue of their empty semantic structures, their interpretation almost always requires the use of antecedent REs.

From the point of view of reference resolution, three main types of relations must thus be distinguished:

- identity coreference: REs that co-specify;
- non-identity coreference: this is a referring relation between two DEs, for instance part/whole, person/function, variable/value, etc.
- anaphora: asymmetric relation between the anaphor and antecedent (with identity coreference or with non-identity one).

In what follows, we will be concerned with *identity coreference between REs*, grouped therefore into DEs.

2 Reference communication: a model

In this section we introduce a conception of reference *communication* between individuals inspired from information theory. Therefore, we introduce a *speaker* (possibly the author of a text) who produces an utterance or linguistic *message* and addresses it to a *hearer* (possibly the reader of a text).

2.1 Referring acts

The notion of *referring act* supposes that:

1. For each utterance, the sender has in mind or *activates* one or several DEs (and one or more properties that concern the DEs).
2. For each DE that the speaker activates in their mind, a fragment of the utterance, specifically related to the activation of that DE, is the RE that specifies the DE.
3. Upon reception, each RE activates one of the receiver's DEs, which may be an existing DE or created on the spot.

It is thus necessary that the REs be understood as such and that the receiver activate a DE upon reception. Significantly, activating a DE for a phrase that did not intend to activate one, or failing to activate a DE where a DE should have been activated is not *misunderstanding* reference, it is *not referring at all.*

2.2 Felicitous referring acts

Intuitively, one would be tempted to say that a referring act was felicitous, or that a *reference was understood*, if the receiver activates the "same" DE as the sender. Unfortunately, this position is difficult to defend, since DEs belonging to different minds cannot be easily compared. Therefore, finding out whether a referring act has been felicitous (i.e., *evaluating reference understanding*) is possible *only* by checking that subsequent referring acts activating the same speaker-DE also activate the same hearer-DE, that is, in terms of *correlations.* So, evaluation must essentially be performed on sets of REs, not on individual referring acts.

2.3 Infelicitous referring acts: r-errors and p-errors

Suppose that after a first referring act, the sender produces a second one, which may activate either the same DE, or another one. The hearer also activates a DE, either the first one, or another one. There are thus four possibilities, two of which are incorrect.

In Figure 1–2a, the sender re-activates the same DE, while the receiver activates another one instead of activating the same one (hence: "the receiver wrongly believes that the sender refers to another object"). So, this *r-error* introduces a rupture in the structure of DEs; it may also be viewed as a "missing link" between two referring acts or two REs.

In Figure 1–2b, the sender activates a different DE on the second referring act, while the receiver re-activates the previous one, instead of switching to another one (hence: "the receiver wrongly believes that the sender refers to the same object"). So, this *p-error* groups together two referring expressions that should not be associated. It may also be viewed as a "wrong link" between two referring acts. Note that a referring act may generate both types of errors simultaneously, provided that at least two referring acts have preceded it.

3 Formal prerequisites to the evaluation of coreference resolution

3.1 Coreference resolution by a computer program

Following the definitions given above, a *coreference resolution program* constructs all the coreference links between REs, whether they are anaphoric or not—most approaches focus only on the *identity coreference*. The transitive closure of all the links generates the RE sets from which the DEs can be abstracted. The goal of such systems is to construct directly the sets of coreferent REs in narrative texts, regardless of their anaphoric relations. Of course, these relations play a central role in the identification of coreference relations.

Anaphora resolution programs, focusing on the asymmetric links between anaphors and their antecedents, do not belong in the previous frame (Barbu and Mitkov 2001, Mitkov 1998).

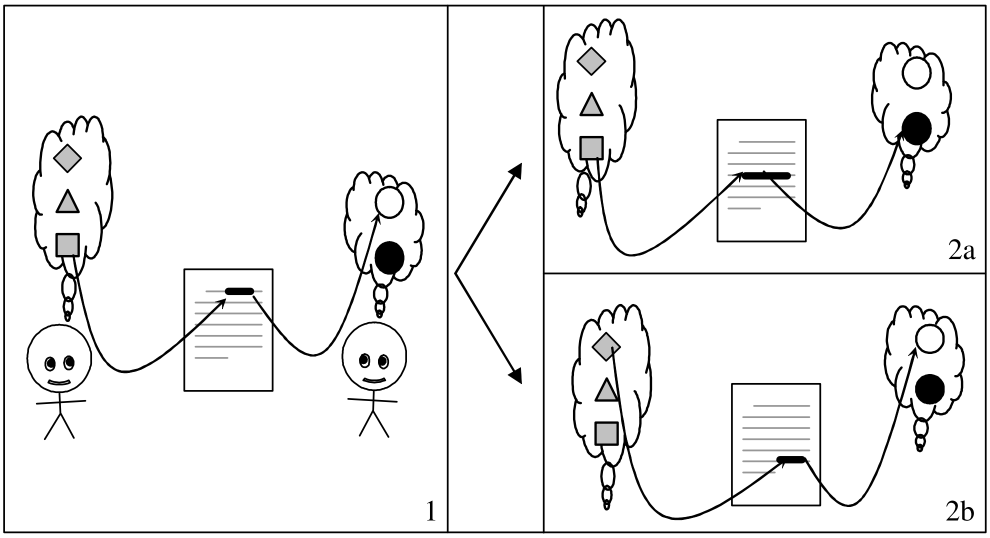

Figure 1. Two types of reference understanding errors: (1) initial referring act, (2a,b) two possible subsequent referring acts – (2a) r-error, (2b) p-error with respect to (1)

3.2 The "equivalence class" view

3.2.1 Intuitive approach

We consider that the set of REs is the same for the sender and the receiver, in order to evaluate specifically reference resolution, as explained above. A program would thus first need a list of correct REs. Then, we consider the distributions of REs into sets of co-specifying REs (activating the same DE), for the speaker (*key*) and for the hearer (*response*).

These sets are *equivalence classes* for the coreference relation, either for the sender or for the receiver (Vilain, Burger, Aberdeen, Connolly and Hirschman 1995). Indeed, an RE belongs to one and only one class (possibly a singleton class) and the classes form a *partition* of the RE set (key partition *vs.* response partition). So, if co-specification links between REs are used, the sets must be built using the transitive closure of the sets of links.

Measuring the proximity between the key and response partitions of the same RE set is not a trivial mathematical problem. Set theory defines only the notion of a partition being more (or less) fine-grained than another one. We will show that information theory provides indirect results on the comparison of partitions.

3.2.2 Definitions

Let E be the set of all REs, and let P_K be the key partition, that is, a set of subsets of E, $P_K = \{K_1, K_2, \ldots, K_n\}$, that are non empty, do not overlap, and recover E (equivalence classes)—see an example in Figure 2 below. Likewise, let the response partition be $P_R = \{R_1, R_2, \ldots, R_m\}$.

The perfect answer corresponds to $P_R = P_K$, so that for each K_i there exists R_j such as $R_j = K_i$. When this is not the case, it is useful to consider all the response classes that contain fragments of a given key class K. The *projection* of K on P_R is first defined as the set of fragments into which K is divided in the response partition:

$$\pi(K) = \{A | \exists\, R_j \in P_R \text{ with } A = K \cap R_j \text{ and } A \neq \varnothing\}$$

The set of response classes that contain these fragments is:

$$\pi^*(K) = \{R_j \mid R_j \in P_R \text{ and } R_j \cap K \neq \varnothing\}$$

Conversely, we define the projection $\sigma(R)$ of a response class R on P_K and the set $\sigma^*(R)$ of key classes containing the fragments.

Since each key class K has at least one projection (itself) and at most $|K|$ (if it is completely fragmented), the following inequalities hold:

(PROP.1) $1 \leq |\pi(K)| \leq |K|$ and $1 \leq |\sigma(R)| \leq |R|$, for all $K \in P_K$ and $R \in P_R$.

In the example on Fig. 2, there are four key classes and three response classes. K_1 and K_2 project onto P_R as single fragments, while K_3 and K_4 are both divided in two—*e.g.*, $\pi(K_3) = \{\{6, 7, 8, 9, 10\}, \{11, 12\}\}$. So, $\pi^*(K_1) = = \{R_1\}$, $\pi^*(K_2) = \{R_2\}$, $\pi^*(K_3) = \{R_1, R_2\}$ and

$\pi^*(K_4) = \{R_2, R_3\}$. Conversely, $\sigma^*(R_1) = \{K_1, K_3\}$, $\sigma^*(R_2) = \{K_2, K_3, K_4\}$, and $\sigma^*(R_3) = \{K_4\}$.

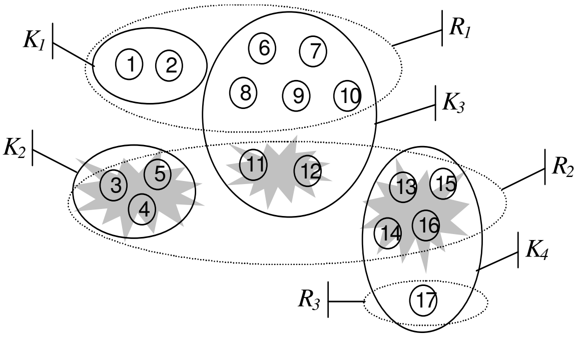

Figure 2. Example of key and response classes (solid *vs.* dashed line), REs being circled numbers. Shaded areas represent $\sigma^*(R_2)$, the projection of R_2 onto P_K.

3.3 The "random variable" view

To apply information theory, we view reference understanding as the constant co-activation of the same DEs for the sender and the receiver. This is grounded in the model of a communication channel (Ash 1965, Cover and Thomas 1991, Shannon and Weaver 1949).

In this model, the sender or *source* is a random variable K that may take several values, and the *receiver* is another random variable R, with values from another set. The "quality" of the communication channel—its *accuracy* and *noiselessness*—is measured using the statistical correlation of K and R.

Here, K is the DE activated by the speaker for each referring act, and R is the DE activated by the hearer upon reception. The probability laws of these random variables are given by the partitions of the set of all referring acts into coreference sets; this ensures that the definitions given below still apply. However, the present model offers a specific method to compare partitions P_K et P_R.

4 An information-theoretic measure of reference transmission accuracy

The measure we propose relies on the concept of *referring information*. The use of the communication channel model allows us to define *mutual referring information* between the speaker and the hearer, and to define "quality" as the maximization of this quantity. This quantity can decrease due to loss or increase due to unjustified gains of referring information. Although based on mutual information (Cover and Thomas 1991), the measure below bears also resemblance with the Kullback-Leibler (1951) divergence.

4.1 Preliminary definitions

The average *referring information* emitted per transmission (referring act), noted $H(P_K)$, is the *entropy* of the source. The average referring information received is $H(P_R)$. Both are defined as follows (E is the set of REs):

$$H(P_K) = -\sum_{K_i \in P_K} \frac{|K_i|}{|E|} \cdot \log \frac{|K_i|}{|E|}$$

$$H(P_R) = -\sum_{R_j \in P_R} \frac{|R_j|}{|E|} \cdot \log \frac{|R_j|}{|E|}$$

The loss of information through the communication channel is defined as the conditioned entropy of the sender given the receiver's value, averaged over these possible values: $H(P_K|P_R)$. Infelicitous referring acts of the *p-error* type increase this loss. Conversely, $H(P_R|P_K)$ measures irrelevant information gained through the channel (this quantity is less used in information theory). Infelicitous referring acts of the *r-error* type increase these gains. Formally $H(P_K|P_R)$ is computed as:

$$-\sum_{(K_i, R_j) \in P_K \times P_R} \frac{|K_i \cap R_j|}{|E|} \cdot \log \frac{|K_i \cap R_j|}{|R_j|}$$

and $H(P_R|P_K)$ is computed as:

$$-\sum_{(K_i, R_j) \in P_K \times P_R} \frac{|K_i \cap R_j|}{|E|} \cdot \log \frac{|K_i \cap R_j|}{|K_i|}$$

with the convention that "$0 \cdot \log(0) = 0$".

Finally, mutual referring information between the speaker and the hearer is defined as: $I(P_K, P_R) = H(P_K) - H(P_K|P_R)$. Proposition (2) below shows that mutual referring information is positive, and that all $H(\ldots)$ values are positive too.

(PROP.2) $0 \leq H(P_R|P_K) \leq H(P_R)$
$0 \leq H(P_K|P_R) \leq H(P_K)$

A well-known result from information theory (Prop. 3) is in our view the fundamental equation of referring *information*, as it reads: "the received referring information equals the information sent, minus the losses, plus the unjustified accruals". This result also shows that mutual information is symmetric $I(P_K, P_R) = I(P_R, P_K)$.

(PROP.3) $H(P_R) = H(P_K) - H(P_K|P_R) + H(P_R|P_K)$

4.2 H-recall and H-precision

We define at this point recall and precision as measures of, respectively, irrelevant information gains (r-errors) and loss of information (p-errors). Since we prefer to normalize these values to the [0;1] interval, we propose the following definitions, where *HRS/HPS* are "recall/precision success":

$$HRS = \frac{H(P_R) - H(P_R|P_K)}{H(P_R)}$$

$$HPS = \frac{H(P_K) - H(P_K|P_R)}{H(P_K)}$$

We must add that $HRS = 1$ when $H(P_R) = 0$, and $HPS = 1$ when $H(P_K) = 0$. The inequalities in (Prop. 2) ensure that *HRS* and *HPS* vary between 0 and 1. Finally, we define f-measure as the harmonic mean between H-recall and H-precision (as usual).

4.3 Properties of the H measure

An important result from information theory ensures that the maximal score is reached when the speaker and the hearer have identical partitions of the RE set, and only in that case:

(PROP.4) $\textit{f-measure} = 100\% \Leftrightarrow$
$H(P_R|P_K) = H(P_K|P_R) = 0 \Leftrightarrow$
$P_R = P_K$

It is also possible to list the cases that yield a 0% *f-measure* score, as in (Prop. 5). The last case is the non trivial one, made explicit in (Prop. 6) further below.

(PROP.5) *f-measure* = 0 *iff* at least one of the following conditions holds:
- $H(P_R) = 0$ & $H(P_K) \neq 0$ (one response class, several key classes)
- $H(P_k) = 0$ & $H(P_R) \neq 0$ (one key class, several response. classes)
- $H(P_k) \neq 0$ & $H(P_R) \neq 0$ & "P_K and P_R are independent"

The last case corresponds to zero mutual referring information between the speaker and the hearer (statistical independence, the speaker understands "nothing" of the hearer's references).

(PROP.6) The following are equivalent:
- "P_K and P_R are independent"
- $H(P_K) = H(P_K|P_R)$
- $H(P_R) = H(P_R|P_K)$
- vectors $(|K_1 \cap R_j|, \ldots, |K_n \cap R_j|)$, $1 \leq j \leq m$, are proportional
- vectors $(|K_i \cap R_1|, \ldots, |K_i \cap R_m|)$, $1 \leq i \leq n$, are proportional

For a given number of response classes $|P_R|$, it is not always possible to arrange response REs so that the score reaches 0%. Of course, a single response class ($|P_R| = 1$) will always get 0% score.

5 Comparison to other measures

We cannot provide in this paper a detailed comparison of this measure to all the others that have been proposed for coreference or anaphora resolution. Given that an automatic evaluation measure as the H one tries to capture the judgment of human experts on the same output, there are no formal arguments that *validate* a measure. We have however defined elsewhere some commonsense criteria for the coherence of an evaluation measure, but we cannot list them here for lack of space.

5.1 Evaluation measures for coreference resolution

Since the first attempt by Vilain *et al.* (1995) to define an evaluation measure for coreference at the MUC-6 conference (MUC-6 1995), several other proposals have tried to analyze and improve existing measures. It is beyond our scope here to analyze each of them in detail. We will summarize their main stance, then compare the scores that they provide. We focus on measures for coreference resolution: the case of anaphora must be dealt with separately (Barbu and Mitkov 2001, Mitkov 2002).

MUC measure (M) – The main contribution of the algorithm proposed by Vilain *et al.* (1995) is a method to count coreference links, which depends only on the sets of coreferent REs (the DEs), not on the particular links that constitute them. The count is indulgent as it computes, by definition, the minimal number of missing and wrong links.

B^3 measure (B) – Bagga and Baldwin (1998), aware of the indulgence of the MUC algorithm, define another recall and precision, *per* RE, then average these values to obtain global scores. Their scores are lower than the MUC scores when many REs are unduly grouped, but they are always well above 0%.

κ-measure (K) – Passonneau (1997) uses the *kappa* factor (Krippendorff 1980) to measure the agreement between two annotators of a given text, based on the probability of agreement by chance (Carletta 1996). For the distance between key and response, the κ-factor is especially relevant when these are very close. The score is computed using MUC recall and precision, and though it is less indulgent, it also bears less information than the MUC couple.

Core-DE measure (C) – The notion of core-DE (Popescu-Belis and Robba 1998) attempted to grasp a program's view (response) of each correct DE. For each key DE, an associated core-DE in the response is computed. Then, all the REs in the key DE that do not belong to the corresponding core-DE count as recall errors. Precision is computed symmetrically. The C-score is lower than the MUC one for every response.

Descriptive specificity measure (D) – Trouilleux *et al.* (2000) attempt to determine the best match between the DE's of the program and the correct ones (the approach is similar to a version of the previous measure called "exclusive-core-DE"). They define an elaborate matching algorithm and scoring function.

On the whole, recall and precision for the M-B-C and H measures vary from 0 to 1. The K-score varies from –1 to +1: +1 for perfect agreement, 0 for random agreement, –1 for "perfect disagreement", *i.e.* for negative statistical correlation between the links.

Each measure has its own advantages and drawbacks, one of the most frequent problems being "indulgence", that is, rather high scores for rather poor answers, or even a minimal score well above zero. Most of the measures are too indulgent when the key has a high RE-to-DE ratio, even a poor answer receiving then a decent score. None of the measures satisfies all our coherence criteria mentioned above. A sound use of these measures consists in comparing the quality of two responses and to determine the best one. Concordant variation of all measures between two responses is a sign of reliability.

5.2 Examples

An empirical method to judge the relevance of evaluation measures is to compare their scores on sample key/response sets, and to find out which measure best reflects "quality".

5.2.1 Synthetic keys and responses

We first use some artificial examples to illustrate the measures on particular cases. These are also handled by our implementation of the measures above. The following examples were tested:

1. The sample text in Figure 2, with 17 REs and the key and response from the figure.
2. A text with ten REs and two key classes, $K_1 = \{1, 2, 3, 4, 5\}$ and $K_2 = \{6, 7, 8, 9, 10\}$. The response is a "nothing was done" response, *i.e.*, $P_R = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}, \{10\}\}$.
3. Using the same text as (2), we suppose now that the response has all REs grouped into one class $R_1 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, hence $P_R = \{R_1\}$.
4. We use now the sample text from the MUC-6 proceedings. There are 147 REs in 15 key classes (singletons are not counted). First, we consider the "nothing was done" response, with $|P_R| = 147$.
5. Using the same text, we suppose now that the response groups all REs into one class, so that $|P_R| = 1$.
6. The same text contains in fact 50 pronouns, but only five key classes contain pronouns. We suppose now that the system is unable to understand pronouns, so it groups into one separate response class, but correctly solves the 97 other REs. So, $|P_R| = 15 + 1$.

Results of the measures are given below, with a three-letter code: M, B, C, or H for the name of the measure, R or P for recall/precision, and S for success.

	MRS	*MPS*	*BRS*	*BPS*	*CRS*	*CPS*	*HRS*	*HPS*
1	85	79	74	49	77	50	55	37
2	0	100	20	100	0	100	30	100
3	100	89	100	50	100	44	100	0
4	0	100	10	100	0	100	40	100
5	100	90	100	19	100	31	100	0
6	96	97	65	79	67	82	76	81

Table 1. Recall and precision for the six sample texts (%)

Example	MUC	B^3	κ	C	H
1	81	59	–18	61	44
2	0	33	0	0	46
3	94	67	0	62	0
4	0	19	0	0	57
5	95	33	0	47	0
6	97	71	66	73	78

Table 2. *F-measures* (in %) and *kappa* for the designed examples

Example (1) obtains relatively high scores (except for κ) despite a confuse response. Examples (2) and (4) show the scores of a system that "performs no resolution", whereas examples (3) and (5) are those of a system that "groups all REs". Both strategies are extremely crude, but we see they sometimes receive high scores (of course, precision is 100% in the first case, and recall is 100% in the second case). The *f-measures* in examples (3) and (5) are quite high, especially for the MUC measure, which clearly proves to be too indulgent. Example (6) is still more realistic, and shows again the indulgence of some measures, in a case when 30% of the REs (the pronouns) are incorrectly resolved.

Regarding relative indulgence, these results confirm that the only comparable measures are C and MUC, the former being less indulgent. The scores are also covariant, that is, increase or decrease simultaneously between two examples. For instance, (6) receives better *f-measure* scores than (4) and (5), for all the measures.

5.2.2 Results of our system on real texts

We have developed a reference processing system, and applied to it the above measures. An annotation module provides the means to define a key partition of the RE set. Three narrative texts were encoded: they are quite long (ca. 100 pages for T3) and have important coreference rates ($|E|$ / $|P_K|$).

Characteristic	T1	T2	T3
Words	2630	7405	28576
REs ($|E|$)	638	686	3359
Key DEs ($|P_K|$)	372	216	480
Coref. rate ($|E|$ / $|P_K|$)	1.72	3.18	7.00
Noun phrase REs	510	390	1864
Pronoun REs	102	262	1398

Table 3. Characteristics of the trial texts

Our program's results (cf. Tables 4 and 5) may seem quite high when compared to programs from the MUC campaign, which scored in the 60% range. In fact, there are two differences in evaluation: we do not evaluate RE identification, giving the system the correct REs from the start, and we use much longer texts with larger key classes, thus biasing the MUC measure towards higher scores.

	MRS	*MPS*	*BRS*	*BPS*	*CRS*	*CPS*	*HRS*	*HPS*
T1	70	78	75	75	53	47	89	89
T2	62	77	50	57	43	36	71	71
T3	70	88	37	52	43	44	59	64

Table 4. System's results on trial texts (in %)

	MUC	B^3	κ	C	H
T1	74	75	57	50	89
T2	69	53	20	39	71
T3	78	43	9	43	61

Table 5. *F-measures* (in %) and κ for the system's results on trial texts (from **Table**)

Despite the similar nature of the three texts, the scores of the program are quite variable. Indeed, the MUC measure, and to a lesser extent the C measure, are more indulgent as the number of REs increases (T1 *vs.* T2 and T2 *vs.* T3), while the system's "quality" is definitely constant. The B^3 and H measures vary in the opposite direction. So, *f-measures* increase for MUC and C, and decrease for B^3 and H when comparing T2 and T3, because they do not have the same bias with respect to the number of REs. However, when applied to texts of similar lengths, all the measures agree in designating the response on T1 as better than the one on T2, reflecting the capacities of the program on those texts.

Conclusion

The present work has presented a measure for coreference resolution based on an information-theoretic model that provides a certain number of theoretical results. The main point has been the idea of "referring information", and the way it can be lost or irrelevantly enriched in the process of communication. The resulting measure provides a relevant score for coreference resolution evaluation.

Many refinements should be considered as future work: the case when speaker and hearer have different RE sets, the problem of non-identity coreference (links between DEs), the problem of anaphora (asymmetric links between an RE and a DE). For lack of space, we could not describe in this paper our proposals for each of these particular challenges.

References

Ash R. B. (1965): Information Theory, New York, NY, Interscience Publishers (John Wiley and Sons).

Bagga A. and Baldwin B. (1998): "Algorithms for Scoring Coreference Chains", Proceedings LREC'98 Workshop on Linguistic Coreference, Granada, Spain.

Barbu C. and Mitkov R. (2001): "Evaluation Tool for Rule-based Anaphora Resolution Methods", Proceedings 39th Annual Meeting of the ACL and 10th Conference of the European Chapter, Toulouse, France, pp. 34-41.

Carletta J. (1996): "Assessing Agreement on Classification Tasks: The Kappa Statistic", Computational Linguistics, vol. 22, n° 2, pp. 249-254.

Cover T. M. and Thomas J. A. (1991): Elements of Information Theory, New York, NY, John Wiley and Sons.

Cristea D., Postolache O.-D., Dima G.-E. and Barbu C. (2002): "AR-Engine - a framework for unrestricted coreference resolution", Proceedings LREC 2002 (Third International Conference on Language Resources and Evaluation), Las Palmas, Canary Islands, Spain, pp. 2000-2007.

Grosz B. J., Joshi A. K. and Weinstein S. (1995): "Centering: A Framework for Modeling the Local Coherence of Discourse", Computational Linguistics, vol. 21, n° 2, pp. 203-225.

Krippendorff K. (1980): Content Analysis: An Introduction to Its Methodology, Beverly Hills, CA, Sage Publications.

Kullback S. and Leibler R. A. (1951): "On information and sufficiency", Ann. Math. Stat., vol. 22, pp. 79-86.

Mitkov R. (1998): "Robust pronoun resolution with limited knowledge", Proceedings COLING-ACL '98, Montréal, Québec, Canada, vol. II/II, pp. 869-875.

Mitkov R. (2002): Anaphora Resolution, London, UK, Longman.

MUC-6 (1995): Proceedings of the 6th Message Understanding Conference (DARPA MUC-6 '95), San Francisco, CA, Morgan Kaufman.

Passonneau R. J. (1997): Applying Reliability Metrics to Co-Reference Annotation, Technical Report Columbia University - Department of Computer Science, CUCS-017-97.

Popescu-Belis A. and Robba I. (1998): "Three New Methods for Evaluating Reference Resolution", Proceedings LREC'98 Workshop on Linguistic Coreference, Granada, Spain.

Shannon C. E. and Weaver W. (1949): The Mathematical Theory of Communication, Urbana, Ill., University of Illinois Press.

Sidner C. L. (1983): "Focusing in the Comprehension of Definite Anaphora", in M. Brady et R. Berwick (ed.), Computational Models of Discourse, Cambridge, MA, MIT Press, pp. 267-330.

Trouilleux F., Gaussier E., Bès G. G. and Zaenen A. (2000): "Coreference Resolution Evaluation Based on Descriptive Specificity", Proceedings LREC 2000 (Second International Conference on Language Resources and Evaluation), Athens, Greece, vol. III/III, pp. 1315-1322.

van Deemter K. and Kibble R. (2000): "On Coreferring: Coreference in MUC and Related Annotation Schemes", Computational Linguistics, vol. 26, n° 4, pp. 629-637.

Vieira R. and Poesio M. (2000): "An Empirically Based System for Processing Definite Descriptions", Computational Linguistics, vol. 26, n° 4, pp. 539-593.

Vilain M., Burger J., Aberdeen J., Connolly D. and Hirschman L. (1995): "A Model-Theoretic Coreference Scoring Scheme", Proceedings 6th Message Understanding Conference (MUC-6), Columbia, Maryland, pp. 45-52.