Coreference Resolution Evaluation Based on Descriptive Specificity

François Trouilleux[*†], Éric Gaussier[*], Gabriel G. Bès[†], Annie Zaenen[*]

[*] Xerox Research Centre Europe
6, Chemin de Maupertuis. 38240 Meylan, France
Firstname.Lastname@xrce.xerox.com

[†] Groupe de recherche dans les industries de la langue (GRIL)
UFR-LACC, Université Blaise-Pascal, Clermont 2
34, avenue Carnot. 63037 Clermont-Ferrand - France
Firstname.Lastname@univ-bpclermont.fr

Abstract

This paper introduces a new evaluation method for the coreference resolution task. Considering that coreference resolution is a matter of linking expressions to discourse referents, we set our evaluation criteron in terms of an evaluation of the denotations assigned to the expressions. This criterion requires that the coreference chains identified in one annotation stand in a one-to-one correspondence with the coreference chains in the other. To determine this correspondence and with a view to keep closer to what human interpretation of the coreference chains would be, we take into account the fact that, in a coreference chain, some expressions are more specific to their referent than others. With this observation in mind, we measure the similarity between the chains in one annotation and the chains in the other, and then compute the optimal similarity between the two annotations. Evaluation then consists in checking whether the denotations assigned to the expressions are correct or not. New measures to analyse errors are also introduced. A comparison with other methods is given at the end of the paper.

Identifying expressions which, in a text, denote the same discourse referent is usually considered a key process in automatic information extraction. However, the question of how to evaluate coreference resolution systems has sometimes been an issue: after the publication by Vilain et al. (1995) of a new evaluation scoring scheme for the Message Understanding Conferences, Popescu-Bellis et al. (1998) and Bagga et al. (1998) each proposed new evaluation methods. In this paper, we, in turn, propose a new evaluation method for the coreference resolution task.

A coreference chain is defined by the property the expressions it contains have to denote a specific discourse referent (1). Our evaluation method so aims at evaluating coreference resolution with respect to this property, the problem being to evaluate whether the discourse referents associated with the expressions are the correct ones (2). From this setting of coreference resolution evaluation in terms of denotation assignment, one derives some constraints on the way two annotations should correspond; in particular, we observe that the fact that some expressions in a coreference chain are more specific to their referent than others has to be taken into account (3). The implementation of our evaluation method meets our requirements by computing the optimal similarity between the coreference chains in two annotations using a linear combination of Dice coefficients over some subsets of coreference chains (4). The recall and precision measures then express a comparison of two sets of denotation assignments. Three complementary measures for errors analysis are also proposed (5). Finally, we show how our evaluation method relates with existing ones (6).

1. Definitions

We call "referential chains" the sets of expressions which, in a text, denote the same discourse referent. Given a text T, the relation between referential chains and discourse referents is such that for each referential chain RC, there exists a unique discourse referent DR, such that:

$$RC = \{x | x \text{ is an expression denoting } DR \text{ in } T\}$$

A referential chain may be a singleton. In the sentence *John loves Mary*, for instance, the set $\{John\}$ (the set of expressions denoting "John" in this text) is a referential chain.

Given a text, the coreference resolution task consists in identifying the referential chains which contains at least two elements. We call such sets "coreference chains".

Let us use as a metalanguage to designate discourse referents a system of index of the form oi, with i a different number for different discourse referents.[1] In the following text, expressions which belong to a coreference chain are surrounded with square brackets and followed by an index which represents the discourse referent denoted by the expression.

> During a joint news conference with Mandela, [Clinton]$_{o1}$ defended [his]$_{o1}$ decision not to make a direct apology to African Americans for [slavery]$_{o2}$, even though [he]$_{o1}$ came close to apologizing to Africans for [it]$_{o2}$.

Figure 1: Example text.

This text contains two coreference chains. Let us use sequences of the form A_{o_i}, with A an upper-case letter, to refer to the referential chain denoting the discourse referent o_i. For our example text, let K_{o1} and K_{o2} be the coref-

[1] The string oi stands for "object i", *i.e.* a particular object in the real or possible world denoted by the text.

erence chains containing the expressions which denote the discourse referents $o1$ and $o2$, respectively.[2]

$$K_{o1} = \{Clinton, his, he\}$$
$$K_{o2} = \{slavery, it\}$$

Incidentally, one may also remark that the example text also contains a number of singleton referential chains; the set $\{Mandela\}$, for instance, is one of them.

2. Evaluation criterion

It is important to note that when dealing with coreference chains we are concerned with "the relationship which holds between a text and the world it denotes".[3] As it appears in our definition of a referential chain, what characterizes the expressions in a coreference chain is their property to denote a specific discourse referent. For instance, the two coreference chains in our example text are characterized by the following five statements:

- $Clinton$ denotes $o1$
- his denotes $o1$
- he denotes $o1$
- $slavery$ denotes $o2$
- it denotes $o2$

Considering that the crucial point in coreference resolution is the property that expressions have to denote a specific referent, we propose an evaluation scheme which will aim at evaluating whether expressions have been given the correct denotation or not. The problem will be, for every expression e in a coreference chain, to find out whether the statement "e denotes the discourse referent oi" is correct. We will call a statement of this form a "denotation assignment".

Mapping expressions to discourse referents (*i.e.* assigning a denotation to them) is, we believe, the expected result of the whole coreference resolution process, from which one expects to be able to collect the information given about each discourse referent in various pieces of the text.

3. Correspondence between key and response

In the general case, the practical issue of evaluation in linguistics is the comparison of two annotations of a text. One of the two annotations is considered to be correct (the "key"); the correctness of the other annotation (the "response") is to be evaluated against the key.

In our case, key and response each contain a set of denotation assignments. We will say that a denotation assignment "e_1 denotes $o1$" in the response is correct if the key contains the denotation assignment "e_1 denotes $o2$" and $o1$ and $o2$ point to the same discourse referent. The problem, now, is to say whether $o1$ and $o2$ are the same: given a set of discourse referents in the key and a set of discourse referents in the response, we have to tell which discourse referent of the key corresponds to which discourse referent of the response, and vice-versa.

[2] K stands for "key", as we will consider this interpretation of the text as the key against which a response R will be evaluated.

[3] We are paraphrasing Dowty (1981).

In order to figure out what the correspondence between the discourse referents of the key and the discourse referents of the response is, we will take advantage of the fact that discourse referents originate in discourse itself. However, before we come to that point, let us set a constraint on the correspondence.

3.1. One-to-one correspondence

From the setting of our evaluation criterion in terms of an evaluation of denotation assignments, one may derive the following observation:

> **Observation 1.** Given e_1 and e_2 two expressions in a text T and given A and B two coreference chain annotations of T, if e_1 and e_2 belong to the same referential chain (*i.e.* a coreference chain) in one annotation and e_1 and e_2 belong to two distinct referential chains (of any kind) in the other, then e_1 and e_2 cannot both have been assigned a correct denotation in the two annotations.

Let us suppose annotation A contains the following coreference chain (with integers referring to expressions):

$$A_{o1} = \{1, 2, 3, 4, 5, 6, 7\}$$

and the expressions in A_{o1} appear to belong to three distinct chains in annotation B:

$$B_{o1'} = \{1, 2\}$$
$$B_{o2'} = \{3, 4\}$$
$$B_{o3'} = \{5, 6, 7\}$$

If the key annotation is A, then the seven expressions have the common property of denoting a specific discourse referent. Among the denotation assignments made in annotation B, we have:

- 1 denotes $o1'$
- 3 denotes $o2'$
- 5 denotes $o3'$

If one considers one of these statements to be correct, then one has to consider the other two to be wrong. More generally, in this case, the denotation assignments can be correct for the expressions of one and only one coreference chain in B, *i.e.* only one of the three discourse referents $o1'$, $o2'$, $o3'$ may correspond to $o1$.

Conversely, if B is the key, the denotation assignments for the expressions in A_{o1} can only be correct for the expressions in one and only one of the three subsets of A_{o1} corresponding respectively to $B_{o1'}$, $B_{o2'}$, $B_{o3'}$, *i.e.* either $o1$ corresponds to $o1'$, or to $o2'$, or to $o3'$, exclusively.

In other words, we consider, as do Pospecu-Bellis et al. (1998) in their "Exclusive Core-MRs" method and contrary to Vilain et al. (1995), that, in the case when A is the key, the system has identified three referents when it should have identified one and that the expressions of two of the three coreference chains have not been assigned the correct denotation. In the case when B is the key, the system has identified only one discourse referent when it should have identified three and the expressions which should have been linked to the two unidentified referents have not been assigned the correct denotation.

To sum up, our evaluation criterion requires a one-to-one correspondence between the discourse referents in the key and the discourse referents in the response.

3.2. Descriptive specificity

Up to now, we have been talking about discourse referents as if we did have direct access to these hypothetical entities in our two annotations, but discourse referents only exist insofar as the human annotator says they exist. Given a text, the human annotator associates expressions to discourse referents and thus identifies referential chains; but we are left with only the result of the process, namely sets of expressions. However, one may consider that the result itself may in turn be interpreted by the human annotator, *i.e.* given coreference chains, a human being will be able to associate discourse referents to them. As interpreting the result of a process is precisely the goal of evaluation, we have to consider how this result would be interpreted by a human interpretor.

Suppose a human observator is given a coreference chain and asked to tell what this referent is, for instance:

$$A_{o1} = \{US\ President\ Bill\ Clinton, his, \\ the\ president, he\}$$

This observator will unambiguously recognize $o1$, describing it as something like "Bill Clinton, the man who is currently president of the United States".

From this example, one may observe that the different expressions in a coreference chain contribute in different ways to the identification of the referent. The expression *US President Bill Clinton* in itself would have been enough to identify the referent, while, on the contrary, given just the expressions *the president*, *his* and *he*, identifying the referent is impossible: in order to be interpreted, these three expressions need to be related to a context by an anaphoric relation.

As another example, let us assume that some coreference resolution algorithm identifies the following coreference chain in the Figure 1 example text:

$$R_{o1'} = \{Mandela, his, he\}$$

Given this chain, one would say that these expressions denote "Mandela", whereas the expressions in K_{o1} denote "Clinton". As a consequence, we will say that the two pronouns his and he have not been assigned the correct denotation.

It must be noted that we do not consider that *Mandela* has been assigned an incorrect denotation. Rather, we would say that $R_{o1'}$ corresponds to the singleton referential chain $\{Mandela\}$ in the key. In order to associate a discourse referent to a referential chain, we need to interpret at least one expression of the chain. It is clear that, for that expression, the denotation assignment is trivial and it follows that for a referential chain RC of cardinality x, there are $x - 1$ denotation assignments to evaluate. This number corresponds to the minimal number of "coreference links" which are needed to build the referential chain.[4]

Being more general, we consider that expressions in a coreference chain may be – at least partially – organized into a hierarchy depending on their descriptive specificity with respect to the discourse referent they denote. As a rule, the most specific expressions in a coreference chain will be the ones which will allow the identification of the discourse referent denoted by the chain. So, with a view to get closer to what a human interpretation of the result would be, we require that coreference resolution evaluation take into account the descriptive specificity of expressions with respect to their referents when it comes to building the correspondence between the discourse referents of the key and those of the response.

4. Computing the correspondence

Our method for comparing the coreference chains of two annotations is divided up into two main steps: in the first step, we look for the optimal (in a sense described below) correspondence between the coreference chains in one annotation and the coreference chains in the other. The correspondence we are looking for is, we insist, the one which will allow us to consider that two coreference chains in correspondence may be safely interpreted as denoting the same discourse referent – from which one will be able to evaluate the denotation assignments proposed in the response against the ones given in the key.

4.1. Similarity between Coreference Chains

To establish an optimal correspondence between coreference chains in two annotations, we need to define a similarity measure between them. Following observation 1 above, such a similarity should be based, even though partially, on the number of expressions the coreference chains have in common. However, since all the expressions in a coreference chain do not contribute in the same way to the identification of the referent, we introduce a hierarchy on the expressions according to their descriptive specificity by partitioning each coreference chain S into subsets. In this article, and for the sake of simplicity, we assume that the task is restricted to coreference between noun phrases, and consider a partition into three subsets, namely, from the most to the less specific, the set of proper names, *PN(S)*, the set of noun phrases with a lexical head, *NP(S)*, and the set of pronouns and possessives, *PRO(S)*. However, our approach and our system are not restricted to specific expressions and partitions.[5]

Given two coreference chains, A_{o_i} and B_{o_j}, we require our similarity measure to comply to the following conditions:

i. if A_{o_i} and B_{o_j} are identical, then their similarity is 1. If they have no element in common, then their similarity is 0

ii. the similarity between A_{o_i} and B_{o_j} should be normalized by the length (*i.e.* the number of expressions) of A_{o_i} and B_{o_j}

[4] Note that in the case of a singleton referential chain, this number is 0. There is nothing to evaluate.

[5] Actually, up to five categories may be used in the current implementation of our system.

iii. the similarity between A_{o_i} and B_{o_j} should be primarily based on the similarity between *PN(*A_{o_i}*)* and *PN(*B_{o_j}*)*, then on the similarity between *NP(*A_{o_i}*)* and *NP(*B_{o_j}*)*, and finally on the similarity between *PRO(*A_{o_i}*)* and *PRO(*B_{o_j}*)*

Condition (i) is a standard requirement for a similarity measure. Condition (ii) is often used in the design of a such a measure. It allows one to avoid giving preference to large, generally noisy, sets over smaller, hopefully more precise, ones. Lastly, condition (iii) reflects the use we want to make of the relative descriptive specificity of expressions.

A simple and widely used measure which meets all these conditions is a linear combination of Dice coefficients computed between *PN(*A_{o_i}*)* and *PN(*B_{o_j}*)*, *NP(*A_{o_i}*)* and *NP(*B_{o_j}*)*, and *PRO(*A_{o_i}*)* and *PRO(*B_{o_j}*)*, and is given by the following formula:

$$sim(A_{o_i}, B_{o_j}) =$$

$$a \times \frac{2 \times |PN(A_{o_i}) \cap PN(B_{o_j})|}{|PN(A_{o_i})| + |PN(B_{o_j})|}$$

$$+b \times \frac{2 \times |NP(A_{o_i}) \cap NP(B_{o_j})|}{|NP(A_{o_i}) + NP(B_{o_j})|}$$

$$+c \times \frac{2 \times |PRO(A_{o_i}) \cap PRO(B_{o_j})|}{|PRO(A_{o_i}) + PRO(B_{o_j})|}$$

where $|\mathcal{A}|$ denotes the cardinal number of $\mathcal{A}$, and a, b, and c are weights satisfying the constraints: $a + b + c = 1$ and $a > b > c$, thus ensuring that all conditions are verified. Furthermore, in order to have a strong reading of condition (iii), we also impose: $a > b + c$.

To set the values for a, b and c, we built a test set with observed and manually built examples, the latter so as to see the behavior of our measure on extreme cases. We then arbitrarily chose: $a = 0.6$, $b = 0.3$ and $c = 0.1$, which led us to the expected results. Other choices are possible, but we believe that, on real examples, any setting for a, b and c within the space defined by the constraints, provided the values for a, b and c are not too close to each other, should lead to the same results.

The reader may have noticed an hidden assumption in the above formula, namely that we know the type (*PN*, *NP*, *PRO*) of the expressions. This knowledge can confidently be provided by lexical look-ups in dictionaries and named entity recognizers, and we assume that we dispose of this information at least for the key. If this information is not provided for the response, we use a variant of the above formula, replacing *PN(R)*, *NP(R)* and *PRO(R)* with *R*.

4.2. Correspondence between Coreference Chains

Once the similarities between coreference chains have been computed, we search for the optimal correspondence, *i.e.* the correspondence which maximizes the overall similarity between the two annotations. Since we are interested in a one-to-one correspondence between coreference chains, we are looking for the correspondence $\mathcal{C}$ verifying:

$$\max_{\mathcal{C}} \sum_{(A_{o_i}, B_{o_j}) \in \mathcal{C}} sim(A_{o_i}, B_{o_j})$$

Several algorithms can be used to find the maximal correspondence or an approximation of it. A widely used heuristic consists in sorting the pair (A_{o_i}, B_{o_j}) in decreasing order of their similarity score, and iteratively selecting the best pair, adding it to the correspondence and removing from the list of remaining pairs the ones which contain one of the elements of the selected best pair. If the annotations contain relations between discourse referents, such as *part of*, *member of*, we can refine the strategy by selecting, in case of equal similarity scores, the pair for which related discourse referents have already been aligned, thus ensuring a better coherence in the set of correspondences.

The above optimization problem can anyway be formulated as a bipartite weighted matching problem, see for example (Ahuja et al., 1993), and the optimal correspondence can be found in this framework. Several options are thus at our disposal here, which call for some remarks:

- we have observed no difference between the solution given by the heuristic and the expected solution on our test set
- the solution with bipartite graphs requires a slight modification of the similarity measure between coreference chains, if we want to make use of relations between discourse referents
- some cases we encountered suggest that we may want to break the one-to-one correspondence we imposed, and rather look for correspondences at different levels of granularity, by considering, for example, groups of coreference chains in addition to coreference chains themselves. We have envisaged different correspondences, using flow networks, an extension of bipartite graphs. However, we do not have strong evidence yet that such an extension is mandatory.

5. Evaluation Measures

Once the optimal correspondence has been established, the denotation assignments of the two annotations can be compared and evaluation measures obtained. In the remainder of the paper, we note $\mathcal{C}(K_{o_i})$ the coreference chain R_{o_j} associated with K_{o_i} in the maximum correspondence (reciprocally, $\mathcal{C}(R_{o_j}) = K_{o_i}$).

In addition to the usual recall and precision measures, we use three measures for error analysis: the "overgeneration", "undergeneration" and "substitution" measures, which we adapted from the measures defined for the MUC Named Entity Task (Chinchor, 1995).

To illustrate how the evaluation measures are determined, we will consider the following response for our text in example 1:

> During a joint news conference with [Mandela]$_{o1'}$, Clinton defended [his]$_{o1'}$ decision not to make a direct apology to [African Americans]$_{o2'}$ for slavery, even though [he]$_{o1'}$ came close to apologizing to [Africans]$_{o2'}$ for it.

Figure 2: Response for example text.

One notes that two coreference chains have been identified:

$$R_{o1'} = \{Mandela, his, he\}$$
$$R_{o2'} = \{AfricanAmericans, Africans\}$$

The correspondence between this response and the key is given in Figure 3 below. For each pair of coreference chains in correspondence, we select an expression belonging to the two sets to represent their common referent. We also select a representative expression for each chain associated with the empty set. All these representative expressions (in *italics*) correspond to the trivial denotation assignment evoked earlier (section 3).

Key		Response
{*Mandela*}	$\Longleftrightarrow$	{*Mandela*, his, he}
{*Clinton*, his, he}	$\Longleftrightarrow$	{*Clinton*}
{*Afr. Americans*}	$\Longleftrightarrow$	{*Afr. Am.*, Africans}
{*slavery*, it}	$\Longleftrightarrow$	{*slavery*}
{*Africans*}	$\Longleftrightarrow$	$\emptyset$
$\emptyset$	$\Longleftrightarrow$	{*it*}

Figure 3: Correspondence example

From this correspondence, one derives the following denotation assignments. In the key:

- *his* denotes "Clinton"
- *he* denotes "Clinton"
- *it* denotes "slavery"

and in the response:

- *his* denotes "Mandela"
- *he* denotes "Mandela"
- *Africans* denotes "African Americans"

leaving aside trivial denotation assignments such as *Mandela* denotes "Mandela".

5.1. Recall and precision

The denominator in the recall measure is the total number of denotation assignments in the key ($possible$). The denominator in the precision measure is the total number of denotation assignments in the response ($actual$). As for a given referential chain A_{o_i}, the number of denotation assignment is $|A_{o_i}| - 1$, we have:

$$possible = \sum_{o_i}(|K_{o_i}| - 1)$$
$$actual = \sum_{o_j}(|R_{o_j}| - 1)$$

It is possible that the maximum correspondence maps some coreference chains in the key and/or in the response to an empty set. The empty set is not a referential chain: both the sum for $possible$ and $actual$ are based only on the discourse referents associated to the referential chains of each of the two annotations in turn.

Recall and precision are then defined in a standard way:

$$recall = \frac{\sum_{o_i}(|K_{o_i} \cap \mathcal{C}(K_{o_i})| - 1)}{\sum_{o_i}(|K_{o_i}| - 1)}$$

$$precision = \frac{\sum_{o_j}(|R_{o_j} \cap \mathcal{C}(R_{o_j})| - 1)}{\sum_{o_j}(|R_{o_j}| - 1)}$$

Even though written differently, the numerators in the recall and precision measures correspond to the same number, namely the number of correct denotation assignments. A denotation assignment DA in the response is correct if it exists in the key. There is no such denotation assignment in our example, so both recall and precision are equal to: 0/3 = 0. This score reflects the idea that the two pronouns *his* and *he* do not refer to Mandela and *African Americans* and *Africans* are not the same people; in all aspects, the response annotation is wrong.

5.2. Substitution, over- and undergeneration

One may want to analyse further the errors in an annotation, which can be done in our system using three measures inspired by the ones developed for the MUC Named Entity Task (Chinchor, 1995): "overgeneration", "undergeneration" and "substitution". To obtain these measures, we count the number of $incorrect$, $spurious$ and $missing$ denotation assignments.

A denotation assignment "e_i denotes o_i" in the response is *incorrect* if there exists in the key a denotation assignment "e_i denotes o_j" and o_j is different from o_i. The expression had to be included in a coreference chain, but it has not been included in the *correct* one. In our example, the denotation assignments

- *his* denotes "Mandela"
- *he* denotes "Mandela"

are incorrect. The two pronouns should have been included in the referential chain denoting "Clinton".

A denotation assignment "e_i denotes o_i" in the response is *spurious* if there is no denotation assignment to e_i in the key. The expression has been taken by the correspondence mechanism as the representative expression of a key referential chain which is mapped to the empty set in the response. To a spurious denotation assignment corresponds the failure to identify a discourse referent. In our example, the denotation assignment in the response

- *Africans* denotes "African Americans"

is spurious. As a consequence of this spurious coreference link, the discourse referent denoted by *Africans* in the key is not identified in the response.

A denotation assignment "e_i denotes o_i" in the key is *missing* in the response if there is no denotation assignment to e_i in the response. The expression has been taken by the correspondence mechanism as the representative expression of a response referential chain which is mapped to the empty set. To a missing denotation assignment corresponds the identification in the response of a discourse referent which does not exist in the key. In our example, the denotation assignment in the key

- *it* denotes "slavery"

is missing in the response. As a consequence of this missing coreference link, the response states that *it* denotes a discourse referent which does not exist in the key.

The sum of incorrect, spurious and missing denotation assignments constitutes the total number of errors E:

$$E = incorrect + spurious + missing$$

This number is the denominator in the substitution, overgeneration and undergeneration measures. The numbers of incorrect, spurious and missing denotation assignments are the numerator in the substitution, overgeneration and undergeneration measures, respectively. For our example, we obtain the following values:

$$substitution = incorrect/E = 2/4 = 0.5$$

$$overgeneration = spurious/E = 1/4 = 0.25$$

$$undergeneration = missing/E = 1/4 = 0.25$$

As a whole, these three measures aim at giving information about the capacity a system has to identify the expressions which should be included in a coreference chain, regardless of *which* coreference chain they should be in. High overgeneration indicates a tendency to include in coreference chains expressions which should not. High undergeneration indicates a tendency not to include in coreference chains expressions which should. High substitution indicates that the expressions which should be included in coreference chains are well identified, but are included in the wrong coreference chains.

6. Discussion

Having detailed our evaluation method for the coreference resolution task, we now compare it with three existing methods: the scoring scheme developed by Vilain et al. (1995) for the MUC-6 coreference task, Popescu-Bellis et al.'s "Exclusive Core MR" method (1998) and Bagga et al.'s B-CUBED algorithm (1998).

6.1. Toy examples

In order to better understand the different evaluation methods, it will be useful to have the scores they produce on some (fictitious) examples: first, our example text (Figure 1) and the corresponding response (Figure 2), then four different response scenarios for the MUC-6 walkthrough article.[6]

The MUC-6 walkthrough article is a Wall Street Journal text for which a key annotation of coreference chains is supplied. The key annotation contains 15 coreference chains with a total of 147 expressions. 50 of these 147 expressions are pronouns; these are spread out into 5 of the 15 coreference chains. We will assume the following four situations:

1. each of the 147 expressions belong to a singleton referential chain in the response (no coreference resolution is done);
2. the 147 expressions are grouped into a unique coreference chain $R_{o1'}$ in the response;
3. the 97 non-pronominal expressions are correctly grouped into 15 coreference chains which correspond to the 15 chains in the key and the 50 pronouns are grouped into an extra 16th coreference chain $R_{o16'}$;
4. the 97 non-pronominal expressions are correctly grouped into 15 coreference chains which correspond to the 15 chains in the key but the system does not attempt to interpret the 50 pronouns, so that each of them belongs to singleton referential chain.

The recall and precision measures output by each method for each of these situations are given in Table 1. Integers in the first column refer to the five situations (0 for our example text, 1 to 4 for the four walkthrough article situations). The next four columns give the recall and precision measures (left and right, respectively) for each method.[7] The last column gives the undergeneration, overgeneration and substitution measures output by our system. In some cases, we are unable to provide the values for XC-MR and B-3. The symbol "–" indicates that the denominator of the precision measure is 0.

6.2. Vilain et al.

The scoring scheme developed by Vilain et al. (1995) for the MUC-6 coreference task is grounded on the idea of coreference links. A coreference chain A_{o_i} is an equivalence set defined by $|A_{o_i}| - 1$ coreference links. The basic idea is to count as errors only the minimal number of links to be added between coreference chains in each of the annotations in order to make them identical. Let us assume we have two coreference chain annotations A and B, containing the following coreference chains, respectively:

$$A_{o_1} = \{1, 2, 3, 4, 5\}$$
$$B_{o_{1'}} = \{1, 2, 3\}$$
$$B_{o_{2'}} = \{4, 5\}$$

The coreference chain A_{o_1} is defined by four coreference links and the two coreference chains $B_{o_{1'}}$ and $B_{o_{2'}}$ by two and one link, respectively. In order to have the two annotations correspond, one would just need to add one coreference link between $B_{o_{1'}}$ and $B_{o_{2'}}$: this missing link constitute the only error according to Vilain et al.'s scheme. This error is a recall error if A is the key and a precision error if B is the key.

The Vilain et al. scoring scheme, in our view, mixes up two different aspects of coreference resolution: the identification of the expressions which should be included in a coreference chain, on the one hand, and the inclusion of these expressions in the proper coreference chains. This is apparent in the difference between the scores produced for situations 3 and 4: when the system groups together the 50 pronouns (3) instead of leaving them as singletons (4), the MUC score significantly increases both in recall and precision, while in our system, only precision is affected.

[6] Some of these scenario have originally been proposed by Popescu-Bellis et al. (1998).

[7] XC-MR refers to the "Exclusive Core-MR" method defined by Popescu-Bellis et al.; DA (for "denotation assignments") refers to our method.

sit.	MUC		XC-MR		B-3		DA		u-g	o-g	sub
0	.33	.33					0	–	.25	.25	.50
1	0	–	.10	1	.10	1	0	–	1	0	0
2	1	.90	.31	.31	1	.19	.27	.24	0	.13	.87
3	.96	.97	.69	.84	.63	.78	.62	.63	.02	0	.98
4	.62	1			.49	1	.62	1	1	0	0

Table 1: Recall and precision scores according to different methods

In other words, the MUC scoring scheme gives some credit to the fact that the expressions which should be included in a coreference chain have been recognized, regardless of what discourse referents the expressions are said to denote. Our evaluation method distinguishes the two aspects: the capacity a system has to recognize the expressions which should be included in coreference chains is captured by the three error analysis measures. As a rule, as texts often contain some large coreference chains, to high values for the substitution measure, will correspond fairly high scores with the MUC method (this is the most manifest in situation 2).

We would argue that the MUC scoring scheme is biased by a focus on the coreference resolution *process* rather than on the coreference resolution *result*. In particular, "the recall (resp. precision) error terms are found by calculating the least number of links that need to be added to the response (resp. the key) in order to have the [coreference chains] align." The evaluation in MUC so appears to be set in terms of "what do I need to do in order to get the correct result?", and not in terms of "is the result I obtain correct or not?". It might very well be the case that on some occasions little would have to be done in order to change the result from quite wrong to fairly correct, but this should be a distinct issue. We believe that the evaluation method we propose allow this distinction: recall and precision analyses the result and three separate measures are used for error analysis.

One may add that our method also avoids a shortcoming of the MUC scoring scheme pointed out by Bagga et al. (1998), namely that this scheme "penalizes the precision numbers equally for all type of errors". As we rely on a one-to-one correspondence, an errant coreference link which group together two large coreference chains will lead to lower scores than an errant link between a large and a small coreference chain.

6.3. Popescu-Bellis et al.

Popescu-Bellis et al. (1998), arguing that the results output by Vilain et al.'s method may be "counterintuitive" in some cases, proposed "three new methods for evaluating reference resolution", among which the second one, called "Exclusive Core-MRs", bears strong similarities with ours. The authors determine a one-to-one correspondence between the coreference chains of the two annotations and count a recall error if an expression belonging to K_{o_i} does not belong to $\mathcal{C}(K_{o_i})$ in the response, and a precision error if an expression belonging to R_{o_j} in the response does not belong to $\mathcal{C}(R_{o_j})$.

The method used by Popescu-Bellis et al. to derive their correspondence can be seen as a special case of our method: by setting $a = b = c = 1$ and removing the denominators in our similarity measure, and by using the heuristic presented above, one arrives at the same correspondence. In this respect, our method presents the advantage of formalizing an optimal correspondence between annotations. Moreover, we make use of descriptive specificity to find the optimal correspondence, thus ensuring that chains in correspondence can be interpreted as denoting the same discourse referent. This property is not necessarily true for other correspondences. For example, given the response proposed for our example text (Figure 2), the method designed by Popescu-Bellis et al. will identify the following correspondence for the chains containing the expressions *Clinton* and *Mandela* in the two annotations:

Key		Response
{Mandela}	$\Longleftrightarrow$	$\emptyset$
{Clinton, his, he}	$\Longleftrightarrow$	{Mandela, his, he}
$\emptyset$	$\Longleftrightarrow$	{Clinton}

Figure 4: Exclusive Core-MR Correspondence

It appears that the "Exclusive Core-MR" method does not guarantee, contrary to our descriptive specificity based method, that the chains in correspondence can be interpreted as denoting the same discourse referent.

One may also note that, while we determine the number of denotation assignments to be evaluated as the cardinal number of a coreference chain minus one, thus discarding trivial denotation assignments, Popescu-Bellis et al. does give some credit to these assignments, so that the recall score in his scheme cannot be 0.[8] This explains the differences between the XC-MR and DA scores in situations 1 and 3 even though both systems yield the same correspondence.

6.4. Bagga *et al.*

In order to correct the shortcoming, mentioned above, in the scoring scheme developed by Vilain et al. (1995), Bagga et al. (1998) proposed a measure, called B-CUBED, which integrates two new ideas:

1. each expression receives a score for recall and precision
2. the overall recall and precision scores are based on a weighted average of scores for each expression and/or each class.

[8] This is also the case in Bagga et al.'s method.

The first point allows to consider an expression within the coreference chain it is placed in as a whole, and thus to normalize its contribution with respect to the length of the chain. This has the effect of differentiating different error types in terms of precision. The second point, on the other hand, is barely mentioned by the authors, who seem to retain a scheme assigning equal weights to each expression and/or coreference chain.

It is certainly possible to make use of the relative descriptive specificity of expressions, within B-CUBED, by assigning them different weights in the computation of the overall recall and precision. Thus, links between noun phrases, proper names and pronouns can be treated differently, and the emphasis can be put on a certain type of expressions for a particular task. However, B-CUBED still aims at evaluating coreference links between expressions, regardless of the discourse referent they denote, and thus suffers from the same weaknesses as Vilain et al.'s scoring scheme.

7. Conclusion

In this paper, we have introduced a new evaluation scheme for coreference resolution, which, rather than setting the problem in terms of linking expressions together, sets it in terms of assigning a denotation to expressions. In order to keep closer to what human interpretation of the results would be, we take into account the descriptive specificity of expressions, thus providing ourselves with an access to discourse referents. Having this access to discourse referents, we are then able to determine the correspondence between two coreference chain annotations with respect to their denotation, from which the correctness of the denotations assigned to expressions may then be evaluated. We would argue that this method, contrary to previously proposed methods, offers a clearer distinction between an evaluation with respect of the expected *result* of coreference resolution (recall and precision) and an evaluation of the coreference resolution process (error analysis measures).

In any case, our proposition may illustrate the fact that when a different light is cast on a particular object, this object might look different.

8. Acknowledgements

Thanks to Pierre Isabelle for very useful comments on this paper. F. Trouilleux' s work is partly supported by a grant from the Association nationale de la recherche technique (CIFRE 050/98).

9. References

Ahuja, R. K., T. L. Magnanti, and J. B. Orlin, 1993. *Network Flows - Theory, Algorithms, and Applications*. Prentice Hall.

Bagga, A. and B. Baldwin, 1998. Algorithms for scoring coreference chains. In *Proceedings of the LREC'98 Workshop on Linguistic Coreference*. Granada, Spain.

Chinchor, N. and G. Dunga, 1995. Four scorers and seven years ago. The scoring method for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. San Francisco: Morgan Kaufmann.

Dowty, D., R. Wall, and S. Peters, 1981. *Introduction to Montague Semantics*. Dordrecht, Holland: D. Reidel Publishing Company.

Grishman, R. and B. Sundheim (eds.), 1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. San Francisco: Morgan Kaufmann.

Hirshman, L. and N. Chinchor, 1998. MUC-7 coreference task definition. version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. http://www.muc.saic.com/: Science Applications International Corporation.

Passoneau, R., 1997. Applying reliability metrics to coreference annotation. Technical Report CUCS-017-97, Columbia University, Department of Computer Science.

Popescu-Bellis, A., 1999a. Évaluation numérique de la résolution de la référence: critiques et propositions. *Traitement Automatique des Langues*, 40(2):117–142.

Popescu-Bellis, A., 1999b. L'évaluation en génie linguistique: un modèle pour vérifier la cohérence des mesures. *Langues: Cahiers d'études et de recherche francophones*, 2:151–162.

Popescu-Bellis, A. and I. Robba, 1998. Three new methods for evaluating reference resolution. In *Proceedings of the LREC'98 Workshop on Linguistic Coreference*. Granada, Spain.

Vilain, M., J. Burger, J. Aberdeen, D. Connolly, and L. Hirshman, 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. San Francisco: Morgan Kaufmann.