Improving Multilingual Summarization: Using Redundancy in the Input to Correct MT errors

Advaith Siddharthan and **Kathleen McKeown**
Columbia University Computer Science Department
1214 Amsterdam Avenue, New York, NY 10027, USA.
{as372,kathy}@cs.columbia.edu

Abstract

In this paper, we use the information redundancy in multilingual input to correct errors in machine translation and thus improve the quality of multilingual summaries. We consider the case of multi-document summarization, where the input documents are in Arabic, and the output summary is in English. Typically, information that makes it to a summary appears in many different lexical-syntactic forms in the input documents. Further, the use of multiple machine translation systems provides yet more redundancy, yielding different ways to realize that information in English. We demonstrate how errors in the machine translations of the input Arabic documents can be corrected by identifying and generating from such redundancy, focusing on noun phrases.

1 Introduction

Multilingual summarization is a relatively nascent research area which has, to date, been addressed through adaptation of existing extractive English document summarizers. Some systems (e.g. SUMMARIST (Hovy and Lin, 1999)) extract sentences from documents in a variety of languages, and translate the resulting summary. Other systems (e.g. Newsblaster (Blair-Goldensohn et al., 2004)) perform translation before sentence extraction. Readability is a major issue for these extractive systems. The output of machine translation software is usually errorful, especially so for language pairs such as Chinese or Arabic and English. The ungrammaticality and inappropriate word choices resulting from the use of MT systems leads to machine summaries that are difficult to read.

Multi-document summarization, however, has information available that was not available during the translation process and which can be used to improve summary quality. A multi-document summarizer is given a set of documents on the same event or topic. This set provides redundancy; for example, each document may refer to the same entity, sometimes in different ways. It is possible that by examining many translations of references to the same entity, a system can gather enough accurate information to improve the translated reference in the summary. Further, as a summary is short and serves as a surrogate for a large set of documents, it is worth investing more resources in its translation; readable summaries can help end users decide which documents they want to spend time deciphering.

Current extractive approaches to summarization are limited in the extent to which they address quality issues when the input is noisy. Some new systems attempt substituting sentences or clauses in the summary with similar text from extraneous but topic related English documents (Blair-Goldensohn et al., 2004). This improves readability, but can only be used in limited circumstances, in order to avoid substituting an English sentence that is not faithful to the original. Evans and McKeown (2005) consider the task of summarizing a mixed data set that contains both English and Arabic news reports. Their approach is to separately summarize information that is contained in only English reports, only Arabic reports, and in both. While the only-English and in-both information can be summarized by selecting text from English reports, the summaries of only-Arabic suffer from the same readability issues.

In this paper, we use principles from information

theory (Shannon, 1948) to address the issue of readability in multilingual summarization. We take as input, multiple machine translations into English of a cluster of news reports in Arabic. This input is characterized by high levels of linguistic noise and by high levels of information redundancy (multiple documents on the same or related topics and multiple translations into English). Our aim is to use automatically acquired knowledge about the English language in conjunction with the information redundancy to perform error correction on the MT. The main benefit of our approach is to make machine summaries of errorful input easier to read and comprehend for end-users.

We focus on noun phrases in this paper. The amount of error correction possible depends on the amount of redundancy in the input and the depth of knowledge about English that we can utilize. We begin by tackling the problem of generating references to people in English summaries of Arabic texts (§2). This special case involves large amounts of redundancy and allows for relatively deep English language modeling, resulting in good error correction. We extend our approach to arbitrary NPs in §3.

The evaluation emphasis in multi-document summarization has been on evaluating content (not readability), using manual (Nenkova and Passonneau, 2004) as well as automatic (Lin and Hovy, 2003) methods. We evaluate readability of the generated noun phrases by computing precision, recall and f-measure of the generated version compared to multiple human models of the same reference, computing these metrics on n-grams. Our results show that our system performs significantly better on precision over two baselines (most frequent initial reference and randomly chosen initial reference). Precision is the most important of these measures as it is important to have a correct reference, even if we don't retain all of the words used in the human models.

2 References to people

2.1 Data

We used data from the DUC 2004 Multilingual summarization task. The Document Understanding Conference (http://duc.nist.gov) has been run annually since 2001 and is the biggest summarization evaluation effort, with participants from all over the world. In 2004, for the first time, there was a multilingual multi-document summarization task. There were 25 sets to be summarized. For each set consisting of 10 Arabic news reports, the participants were provided with 2 different machine translations into English (using translation software from ISI and IBM). The data provided under DUC includes 4 human summaries for each set for evaluation purposes; the human summarizers were provided a human translation into English of each of the Arabic New reports, and did not have to read the MT output that the machine summarizers took as input.

2.2 Task definition

An analysis of premodification in initial references to people in DUC human summaries for the monolingual task from 2001–2004 showed that 71% of premodifying words were either title or role words (eg. *Prime Minister, Physicist* or *Dr.*) or temporal role modifying adjectives such as *former* or *designate*. Country, state, location or organization names constituted 22% of premodifying words. All other kinds of premodifying words, such as *moderate* or *loyal* constitute only 7%. Thus, assuming the same pattern in human summaries for the multilingual task (cf. section 2.6 on evaluation), our task for each person referred to in a document set is to:

1. Collect all references to the person in both translations of each document in the set.
2. Identify the correct roles (including temporal modification) and affiliations for that person, filtering any noise.
3. Generate a reference using the above attributes and the person's name.

2.3 Automatic semantic tagging

As the task definition above suggests, our approach is to identify particular semantic attributes for a person, and generate a reference formally from this semantic input. Our analysis of human summaries tells us that the semantic attributes we need to identify are `role`, `organization`, `country`, `state`, `location` and `temporal modifier`. In addition, we also need to identify the `person name`. We used BBN's IDENTIFINDER (Bikel et al., 1999) to mark up person names, organizations and locations. We marked up countries and (American) states using a list obtained from the CIA factsheet[1].

[1] *http://www.cia.gov/cia/publications/factbook* provides a list of countries and states, abbreviations and adjectival forms, for example *United Kingdom/U.K./British/Briton* and *California/Ca./Californian.*

To mark up roles, we used a list derived from WordNet (Miller et al., 1993) hyponyms of the *person* synset. Our list has 2371 entries including multiword expressions such as *chancellor of the exchequer*, *brother in law*, *senior vice president* etc. The list is quite comprehensive and includes roles from the fields of sports, politics, religion, military, business and many others. We also used WordNet to obtain a list of 58 temporal adjectives. WordNet classifies these as pre- (eg. *occasional*, *former*, *incoming* etc.) or post-nominal (eg. *elect*, *designate*, *emeritus* etc.). This information is used during generation. Further, we identified elementary noun phrases using the LT TTT noun chunker (Grover et al., 2000), and combined `NP of NP` sequences into one complex noun phrase. An example of the output of our semantic tagging module on a portion of machine translated text follows:

> ...<NP> <ROLE> representative </ROLE> of <COUNTRY> Iraq </COUNTRY> of the <ORG> United Nations </ORG> <PERSON> Nizar Hamdoon </PERSON> </NP> that <NP> thousands of people </NP> killed or wounded in <NP> the <TIME> next </TIME> few days four of the aerial bombardment of <COUNTRY> Iraq </COUNTRY> </NP>...

Our principle data structure for this experiment is the attribute value matrix (AVM). For example, we create the following AVM for the reference to Nizar Hamdoon in the tagged example above:

`name`	Nizar Hamdoon
`role`	representative
`country`	Iraq $(arg1)$
`organization`	United Nations $(arg2)$

Note that we store the relative positions (*arg 1* and *arg 2*) of the country and organization attributes. This information is used both for error reduction and for generation as detailed below. We also replace adjectival country attributes with the country name, using the correspondence in the CIA factsheet.

2.4 Identifying redundancy and filtering noise

We perform coreference by comparing AVMs. Because of the noise present in MT (For example, words might be missing, or proper names might be spelled differently by different MT systems), simple name comparison is not sufficient. We form a coreference link between two AVMs if:

1. The last name and (if present) the first name match.
2. OR, if the role, country, organization and time attributes are the same.

The assumption is that in a document set to be summarized (which consists of related news reports), references to people with the same affiliation and role are likely to be references to the same person, even if the names do not match due to spelling errors. Thus we form one AVM for each person, by combining AVMs. For Nizar Hamdoon, to whom there is only one reference in the set (and thus two MT versions), we obtain the AVM:

`name`	Nizar Hamdoon(2)
`role`	representative(2)
`country`	Iraq(2) $(arg1)$
`organization`	United Nations(2) $(arg2)$

where the numbers in brackets represents the counts of this value across all references. The *arg* values now represent the most frequent ordering of these organizations and countries in the input references. As an example of a combined AVM for a person with a lot of references, consider:

`name`	Zeroual(24), Liamine Zeroual(20)
`role`	president(23), leader(2)
`country`	Algeria(18) $(arg1)$
`organization`	Renovation Party(2) $(arg1)$, AFP(1) $(arg1)$
`time`	former(1)

This example displays common problems when generating a reference. Zeroual has two affiliations - Leader of the Renovation Party, and Algerian President. There is additional noise - the values *AFP* and *former* are most likely errors. As none of the organization or country values occur in the same reference, all are marked *arg1*; no relative ordering statistics are derivable from the input. For an example demonstrating noise in spelling, consider:

`name`	Muammar Qaddafi(10), Muammar Gaddafi(10), Qaddafi(4), Gaddafi(4)
`role`	leader colonel(12), colonel(4) leader(3), minister(2), justice(1)
`country`	Libya(7) $(arg1)$
`organization`	Peace Country(2) $(arg2)$, Country Peace(1) $(arg1)$

Our approach to removing noise is to:

1. Select the most frequent name with more than one word (this is the most likely full name).
2. Select the most frequent role.
3. Prune the AVM of values that occur with a frequency below an empirically determined threshold.

Thus we obtain the following AVMs for the three examples above:

name	Nizar Hamdoon
role	representative
country	Iraq ($arg1$)
organization	United Nations ($arg2$)

name	Liamine Zeroual
role	president
country	Algeria ($arg1$)

name	Muammar Qaddafi
role	leader colonel
country	Libya ($arg1$)

This is the input semantics for our generation module described in the next section.

2.5 Generating references from AVMs

In order to generate a reference from the words in an AVM, we need knowledge about syntax. The syntactic frame of a reference to a person is determined by the role. Our approach is to automatically acquire these frames from a corpus of English text. We used the Reuters News corpus for extracting frames. We performed the semantic analysis of the corpus, as in §2.3; syntactic frames were extracted by identifying sequences involving locations, organizations, countries, roles and prepositions. An example of automatically acquired frames with their maximum likelihood probabilities for the role *ambassador* is:

ROLE=ambassador

(p=.35)	COUNTRY ambassador PERSON
(.18)	ambassador PERSON
(.12)	COUNTRY ORG ambassador PERSON
(.12)	COUNTRY ambassador to COUNTRY PERSON
(.06)	ORG ambassador PERSON
(.06)	COUNTRY ambassador to LOCATION PERSON
(.06)	COUNTRY ambassador to ORG PERSON
(.03)	COUNTRY ambassador in LOCATION PERSON
(.03)	ambassador to COUNTRY PERSON

These frames provide us with the required syntactic information to generate from, including word order and choice of preposition. We select the most probable frame that matches the semantic attributes in the AVM. We also use a default set of frames shown below for instances where no automatically acquired frames exist:

ROLE=<Default>

COUNTRY ROLE PERSON
ORG ROLE PERSON
COUNTRY ORG ROLE PERSON
ROLE PERSON

If no frame matches, organizations, countries and locations are dropped one by one in decreasing order of argument number, until a matching frame is found. After a frame is selected, any prenominal temporal adjectives in the AVM are inserted to the left of the frame, and any postnominal temporal adjectives are inserted to the immediate right of the role in the frame. Country names that are not objects of a preposition are replaced by their adjectival forms (using the correspondences in the CIA factsheet). For the AVMs above, our generation module produces the following referring expressions:

- Iraqi United Nations representative Nizar Hamdoon
- Algerian President Liamine Zeroual
- Libyan Leader Colonel Muammar Qaddafi

2.6 Evaluation

To evaluate the referring expressions generated by our program, we used the manual translation of each document provided by DUC. The drawback of using a summarization corpus is that only one human translation is provided for each document, while multiple model references are required for automatic evaluation. We created multiple model references by using the initial references to a person in the manual translation of each input document in the set in which that person was referenced. We calculated unigram, bigram, trigram and fourgram precision, recall and f-measure for our generated references evaluated against multiple models from the manual translations. To illustrate the scoring, consider evaluating a generated phrase *"a b d"* against three model references *"a b c d"*, *"a b c"* and *"b c d"*. The bigram precision is $1/2 = 0.5$ (one out of two bigrams in generated phrase occurs in the model set), bigram recall is $2/7 = 0.286$ (two out of 7 bigrams in the models occurs in the generated phrase) and f-measure ($f = 2p \times r/(p + r)$) is 0.364. For fourgrams, P, R and F are zero, as there is a fourgram in the models, but none in the generated NP.

We used 6 document sets from DUC'04 for development purposes and present the average P, R and F for the remaining 18 sets in Table 1. There were 210 generated references in the 18 testing sets. The table also shows the popular BLEU (Papineni et al., 2002) and NIST[2] MT metrics. We also provide two baselines - most frequent initial reference to the person in the input (Base1) and a randomly selected initial reference to the person (Base2). As Table 1 shows, Base1 performs better than random selection. This

[2]http://www.nist.gov/speech/tests/mt/resources/scoring.htm

UNIGRAMS	P_{av}	R_{av}	F_{av}
Generated	0.847*@	0.786	0.799*@
Base1	0.753*	0.805	0.746*
Base2	0.681	0.767	0.688
BIGRAMS	P_{av}	R_{av}	F_{av}
Generated	0.684*@	0.591	0.615*
Base1	0.598*	0.612	0.562*
Base2	0.492	0.550	0.475
TRIGRAMS	P_{av}	R_{av}	F_{av}
Generated	0.514*@	0.417	0.443*
Base1	0.424*	0.432	0.393*
Base2	0.338	0.359	0.315
FOURGRAMS	P_{av}	R_{av}	F_{av}
Generated	0.411*@	0.336	0.351*
Base1	0.320	0.360*	0.302
Base2	0.252	0.280	0.235

@ Significantly better than Base1
* Significantly better than Base2
(Significance tested using unpaired t-test at 95% confidence)

MT Metrics	Generated	Base1	Base2
BLEU	0.898	0.499	0.400
NIST	8.802	6.423	5.658

Table 1: Evaluation of generated reference

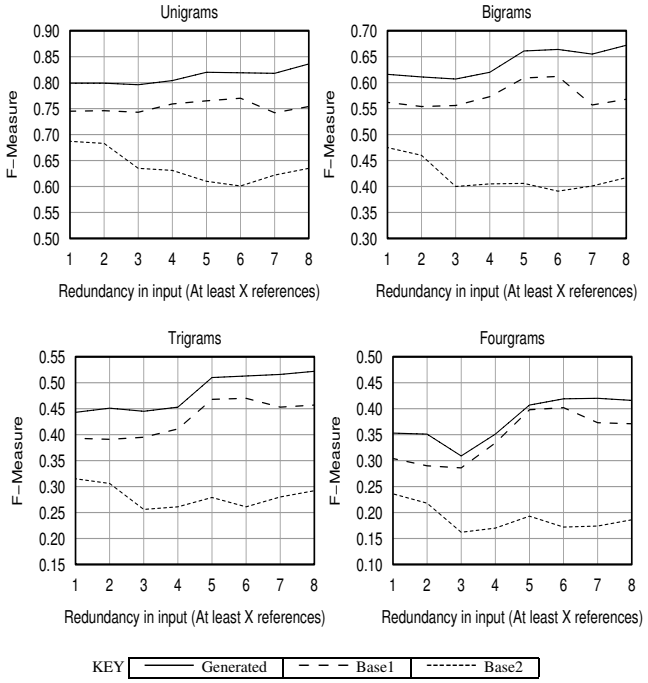

Figure 1: Improvement in F-measure for n-grams in output with increased redundancy in input.

is intuitive as it also uses redundancy to correct errors, at the level of phrases rather than words. The generation module outperforms both baselines, particularly on precision - which for unigrams gives an indication of the correctness of lexical choice, and for higher ngrams gives an indication of grammaticality. The unigram recall of 0.786 indicates that we are not losing too much information at the noise filtering stage. Note that we expect a low R_{av} for our approach, as we only generate particular attributes that are important for a summary. The important measure is P_{av}, on which we do well. This is also reflected in the high scores on BLEU and NIST.

It is instructive to see how these numbers vary as the amount of redundancy increases. Information theory tells us that information should be more recoverable with greater redundancy. Figure 1 plots f-measure against the minimum amount of redundancy. In other words, the value at X=3 gives the f-measure averaged over all people who were mentioned at least thrice in the input. Thus X=1 includes all examples and is the same as Table 1.

As the graphs show, the quality of the generated reference improves appreciably when there are at least 5 references to the person in the input. This is a convenient result for summarization because people who are mentioned more frequently in the input are more likely to be mentioned in the summary.

2.7 Advantages over using extraneous sources

Our approach performs noise reduction and generates a reference from information extracted from the machine translations. Information about a person can be obtained in other ways; for example, from a database, or by collecting references to the person from extraneous English-language reports. There are two drawbacks to using extraneous sources:

1. People usually have multiple possible roles and affiliations, so descriptions obtained from an external source might not be appropriate in the current context.
2. Selecting descriptions from external sources can change perspective — one country's terrorist is another country's freedom fighter.

In contrast, our approach generates references that are appropriate and reflect the perspectives expressed in the source.

3 Arbitrary noun phrases

In the previous section, we showed how accurate references to people can be generated using an information theoretic approach. While this is an important result in itself for multilingual summarization, the same approach can be extended to correct errors in noun phrases that do not refer to people. This extension is trickier to implement, however, because:

1. *Collecting redundancy*: Common noun coreference is a hard problem, even within a single clean English text, and harder still across multiple MT texts.

2. *Generating*: The semantics for an arbitrary noun phrase cannot be defined sufficiently for formal generation; hence our approach is to select the most plausible of the coreferring NPs according to an inferred language model. When sufficient redundancy exists, it is likely that there is at least one option that is superior to most.

Interestingly, the nature of multi-document summarization allows us to perform these two hard tasks. We follow the same theoretical framework (identify redundancy, and then generate from this), but the techniques we use are necessarily different.

3.1 Alignment of NPs across translations

We used the BLAST algorithm (Altschul et al., 1997) for aligning noun phrases between two translations of the same Arabic sentence. We obtained the best results when each translation was analyzed for noun chunks, and the alignment operation was performed over sequences of words and *<NP>* and *</NP>* tags. BLAST is an efficient alignment algorithm that assumes that words in the two sentences are roughly in the same order from a global perspective. As neither of the MT systems used performs much clause or phrase reorganization, this assumption is not a problem for our task. An example of two aligned sentences is shown in figure 2. We then extract coreferring noun phrases by selecting the text between aligned *<NP>* and *</NP>* tags; for example:

1. the Special Commission in charge of disarmament of Iraq's weapons of mass destruction
2. the Special Commission responsible disarmament Iraqi weapons of mass destruction

3.2 Alignment of NPs across documents

This task integrates well with the clustering approach to multi-document summarization (Barzilay, 2003), where sentences in the input documents are first clustered according to their similarity, and then one sentence is generated from each cluster. This clustering approach basically does at the level of sentences what we are attempting at the level of noun phrases. After clustering, all sentences within a cluster should represent similar information. Thus, similar noun phrases in sentences within a cluster are likely to refer to the same entities. We do noun phrase coreference by identifying lexically similar noun phrases within a cluster. We use *SimFinder* (Hatzivassiloglou et al., 1999) for sentence clustering and the f-measure for word overlap to compare noun phrases. We set a threshold for deciding coreference by experimenting on the 6 development sets (cf. §2.6)– the most accurate coreference occurred with a threshold of f=0.6 and a constraint that the two noun phrases must have at least 2 words in common that were neither determiners nor prepositions. For the reference to the *UN Special Commission* in figure 2, we obtained the following choices from alignments and coreference across translations and documents within a sentence cluster:

1. the United nations Special Commission in charge of disarmament of Iraq's weapons of mass destruction
2. the the United Nations Special Commission responsible disarmament Iraqi weapons of mass destruction
3. the Special Commission in charge of disarmament of Iraq's weapons of mass destruction
4. the Special Commission responsible disarmament Iraqi weapons of mass destruction
5. the United nations Special Commission in charge of disarmament of Iraq's weapons of mass destruction
6. the Special Commission of the United Nations responsible disarmament Iraqi weapons of mass destruction

Larger sentence clusters represent information that is repeated more often across input documents; hence the size of a cluster is indicative of the importance of that information, and the summary is composed by considering each sentence cluster in decreasing order of size and generating one sentence from it. From our perspective of fixing errors in noun phrases, there is likely to be more redundancy in a large cluster; hence this approach is likely to work better within clusters that are important for generating the summary.

3.3 Generation of noun phrases

As mentioned earlier, formal generation from a set of coreferring noun phrases is impractical due to the unrestricted nature of the underlying semantics. We thus focus on selecting the best of the possible options — the option with the least garbled word order; for example, selecting 1) from the following:

1. the malicious campaigns in some Western media
2. the campaigns tendentious in some of the media Western European

The basic insight that we utilize is — when two words in a NP occur together in the original documents more often than they should by chance, it is likely they really should occur together in the generated NP. Our approach therefore consists of identifying collocations of length two. Let the number of words in the input documents be N. For each

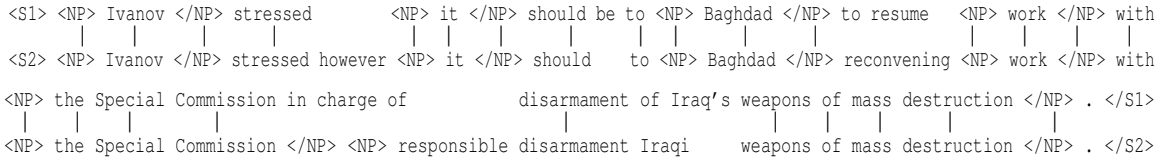

Figure 2: Two noun chunked MT sentences (S1 and S2) with the words aligned using BLAST.

pair of words a and b, we use maximum likelihood to estimate the probabilities of observing the strings "$a\,b$", "a" and "b". The observed frequency of these strings in the corpus divided by the corpus size N gives the maximum likelihood probabilities of these events $p(a, b)$, $p(a)$ and $p(b)$. The natural way to determine how dependent the distributions of a and b are is to calculate their mutual information (Church and Hanks, 1991):

$$I(a, b) = \log_2 \frac{p(a, b)}{p(a) \times p(b)}$$

If the occurrences of a and b were completely independent of each other, we would expect the maximum likelihood probability $p(a, b)$ of the string "$a\ \ b$" to be $p(a) \times p(b)$. Thus mutual information is zero when a and b are independent, and positive otherwise. The greater the value of $I(a, b)$, the more likely that "$a\ \ b$" is a collocation. Returning to our problem of selecting the best NP from a set of coreferring NPs, we compute a score for each NP (consisting of the string of words $w_1...w_n$) by averaging the mutual information for each bigram:

$$Score(w_1...w_n) = \frac{\sum_{i=1}^{i=n-1} I(w_i, w_{i+1})}{n-1}$$

We then select the NP with the highest score. This model successfully selects *the malicious campaigns in some Western media* in the example above and *the United nations Special Commission in charge of disarmament of Iraq's weapons of mass destruction* in the example in §3.2.

3.4 Automatic Evaluation

Our approach to evaluation is similar to that for evaluating references to people. For each collection of coreferring NPs, we identified the corresponding model NPs from the manual translations of the input documents by using the BLAST algorithm for word alignment between the MT sentences and the corresponding manually translated sentence. Table 2 below gives the average unigram, bigram, trigram and fourgram precision, recall and f-measure for the selected NPs, evaluated against the models. We excluded references to people as these were treated formally in §2. This left us with 961 noun phrases from the 18 test sets to evaluate. Table 2 also provides the BLEU and NIST MT evaluation scores.

UNIGRAMS	P_{av}	R_{av}	F_{av}
Mutual information	0.615*@	0.658	0.607*
Base1	0.584	0.662	0.592
Base2	0.583	0.652	0.586
BIGRAMS	P_{av}	R_{av}	F_{av}
Mutual information	0.388*@	0.425*	0.374*@
Base1	0.340	0.402	0.339
Base2	0.339	0.387	0.330
TRIGRAMS	P_{av}	R_{av}	F_{av}
Mutual information	0.221*@	0.204*	0.196*@
Base1	0.177	0.184	0.166
Base2	0.181	0.171	0.160
FOURGRAMS	P_{av}	R_{av}	F_{av}
Mutual information	0.092*	0.090*	0.085*
Base1	0.078	0.080	0.072
Base2	0.065	0.066	0.061

@ Significantly better than Base1
* Significantly better than Base2
(Significance tested using unpaired t-test at 95% confidence)

MT Metrics	Mutual information	Base1	Base2
BLEU	0.276	0.206	0.184
NIST	5.886	4.979	4.680

Table 2: Evaluation of noun phrase selection

We again provide two baselines - most frequent NP in the set (Base1) and a randomly selected NP from the set (Base2). The numbers in Table 2 are lower than those in Table 1. This is because generating references to people is a more restricted problem – there is less error in MT output, and a formal generation module is employed for error reduction. In the case of arbitrary NPs, we only select between the available options. However, the information theoretic approach gives significant improvement for the arbitrary NP case as well, particularly for precision, which is an indicator of grammaticality.

3.5 Manual Evaluation

To evaluate how much impact the rewrites have on summaries, we ran our summarizer on the 18 test sets, and manually evaluated the selected sentences

and their rewritten versions for accuracy and fluency. There were 118 sentences, out of which 94 had at least one modification after the rewrite process. We selected 50 of these 94 sentences at random and asked 2 human judges to rate each sentence and its rewritten form on a scale of 1–5 for accuracy and fluency[3]. We used 4 human judges, each judging 25 sentence pairs. The original and rewritten sentences were presented in random order, so judges did not know which sentences were rewritten. Fluency judgments were made before seeing the human translated sentence, and accuracy judgments were made by comparing with the human translation. The average scores before and after rewrite were 2.08 and 2.26 respectively for fluency and 3.00 and 3.19 respectively for accuracy. Thus the rewrite operations increases both scores by around 0.2.

4 Conclusions and future work

We have demonstrated how the information redundancy in the multilingual multi-document summarization task can be used to reduce MT errors. We do not use any related English news reports for substituting text; hence our approach is not likely to change the perspectives expressed in the original Arabic news to those expressed in English news reports. Further, our approach does not perform any corrections specific to any particular MT system. Thus the techniques described in this paper will remain relevant even with future improvements in MT technology, and will be redundant only when MT is perfect. We have used the Arabic-English data from DUC'04 for this paper, but our approach is equally applicable to other language pairs. Further, our techniques integrate easily with the sentence clustering approach to multi-document summarization – sentence clustering allows us to reliably identify noun phrases that corefer across documents.

In this paper we have considered the case of noun phrases. In the future, we plan to consider other types of constituents, such as correcting errors in verb groups, and in the argument structure of verbs. This will result in a more generative and less extractive approach to summarization - indeed the case for generative approaches to summarization is more convincing when the input is noisy.

[3]We followed the DARPA/LDC guidelines from http://ldc.upenn.edu/Projects/TIDES/Translation/TranAssessSpec.pdf. For fluency, the scale was 5:Flawless, 4:Good, 3:Non-native, 2:Disfluent, 1:Incomprehensible. The accuracy scale for information covered (comparing with human translation) was 5:All, 4:Most, 3:Much, 2:Little, 1:None.

References

S.F. Altschul, T. L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 17(25):3389–3402.

R. Barzilay. 2003. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University, New York.

D. Bikel, R. Schwartz, and R. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34:211–231.

S. Blair-Goldensohn, D. Evans, V. Hatzivassiloglou, K. McKeown, A. Nenkova, R. Passonneau, B. Schiffman, A. Schlajiker, A. Siddharthan, and S. Siegelman. 2004. Columbia University at DUC 2004. In *Proceedings of DUC'04*, pages 23–30, Boston, USA.

K. Church and P. Hanks. 1991. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

D. Evans and K. McKeown. 2005. Identifying similarities and differences across english and arabic news. In *Proceedings of International Conference on Intelligence Analysis*, pages 23–30, McLean, VA.

C. Grover, C. Matheson, A. Mikheev, and M. Moens. 2000. LT TTT - A flexible tokenisation tool. In *Proceedings of LREC'00*, pages 1147–1154.

V. Hatzivassiloglou, J. Klavans, and E. Eskin. 1999. Detecting text similarity over short passages: exploring linguistic feature combinations via machine learning. In *Proceedings of EMNLP'99*, MD, USA.

E.H. Hovy and Chin-Yew Lin. 1999. Automated text summarization in summarist. In I. Mani and M. Maybury, editors, *Advances in Automated Text Summarization*, chapter 8. MIT Press.

C. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL'03*, Edmonton.

G.A. Miller, R. Beckwith, C.D. Fellbaum, D. Gross, and K. Miller. 1993. Five Papers on WordNet. Technical report, Princeton University, Princeton, N.J.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL 2004: Main Proceedings*, pages 145–152, Boston, MA, USA.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL'02*.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Tech. Journal*, 27:379–423.