When is an Embedded MT System "Good Enough" for Filtering?

Clare R. Voss
Army Research Laboratory
Adelphi, MD 20783
voss@arl.mil

Carol Van Ess-Dykema
Department of Defense
Ft. Meade, MD
cjvanes@afterlife.ncsc.mil

Abstract

This paper proposes an end-to-end process analysis template with replicable measures to evaluate the filtering performance of a Scan-OCR-MT system. Preliminary results[1] across three language-specific FALCon[2] systems show that, with one exception, the derived measures consistently yield the same performance ranking: Haitian Creole at the low end, Arabic in the middle, and Spanish at the high end.

1 The Filtering Problem

How do people quickly determine whether a particular foreign language text document is relevant to their interest when they do not understand that foreign language? FALCon, our embedded MT system, has been designed to assist an English-speaking person in filtering, i.e., deciding which foreign language documents are worth having an expert translator process further. In this paper, we seek to determine when such systems are "good enough" for filtering.

We define "filtering" to be a forced-choice decision-making process on individual documents, where each document is assigned a single value, either a "yes, relevant" or a "no, irrelevant" by the system user.[3] The single-document relevance assessment is performed independent of the content of other documents in the processing collection.

When Church and Hovy (1993) introduced the notion that "crummy" MT engines could be put to good use on tasks less-demanding than publication-quality translation, MT research efforts did not typically evaluate system performance in the context of specific tasks. (Sparck Jones and Galliers, 1996). In the last few years, however, the Church and Hovy insight has led to innovative experiments, like those reported by Resnik (1997), Pomarede et al. (1998), and Taylor and White (1998), using task-based evaluation methods. Most recently, research on task-based evaluation has been proposed within TIDES, a recent DARPA initiative whose goals include enabling English-speaking individuals to access, correlate, and interpret multilingual sources of information (DARPA, 1999; Harmon, 1999).

This paper introduces a method of assessing when an embedded MT system is "good enough" for the filtering of hard-copy foreign language (FL) documents by individuals with no knowledge of that language. We describe preliminary work developing measures on *system-internal* components that assess: (i) the flow of words relevant to the filtering task and domain through the steps of document processing in our embedded MT system, and (ii) the level of "noise," i.e., processing errors, passing through the system. We present an *analysis template* that displays the processing steps, the sequence of document versions, and the basic measures of our evaluation method. After tracing the processing of Spanish, Arabic, and Haitian Creole parallel texts that is recorded in the analysis templates, we discuss our preliminary results on the filtering performance of the three language-specific embedded MT systems from this process flow.

---

[1] For a more extensive report of our work, see Voss and Van Ess-Dykema (2000).

[2] FALCon (Forward Area Language CONverter) is a laptop-based embedded MT system integrated at the Army Research Laboratory for field use. (Fisher and Voss, 1997)

[3] See the report entitled "Multilingual Information Management: Current Levels and Future Abilities" for other definitions of filtering, available at http://www.cs.cmu.edu/People/ref/mlim/.

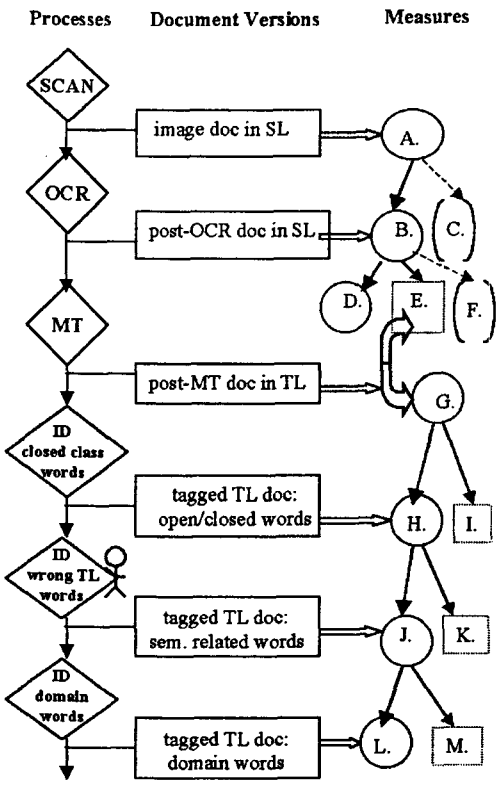

Figure 1 Analysis Template

2 An Embedded MT System Design[4]

Our three systems process documents using a sequence of three software modules. First, the Scan software module creates an online bitmap image in real-time as the user feeds the document into the page-feed scanner.[5] Second, the optical character recognition (OCR) software converts that image to character text and, third, the machine translation (MT) software converts the foreign language character text to English, where it may be stored to disk or displayed on screen directly to the user. The user interface only requires that the user push one or two buttons to carry out all of the system's processing on an individual document.

We tested three separate *language-specific* embedded MT systems for Spanish, Arabic and Haitian Creole. These systems differ in their OCR and MT components, but otherwise they share the same software, Omnipage's Paperport for scaning and Windows95 as the operating system.[6]

3 Approach

As we sought to measure the performance of each component in the systems, it quickly became apparent that not all available measures may be equally applicable for our filtering task. For example, counting the number of source language (SL) *characters* correctly OCR-ed may be overly specific: as discussed below, we only need to make use of the number of SL *words* that are correctly OCR-ed. In the sections to follow, we describe those measures that have been most informative for the task of filtering.

Analysis Template

We use three types of information in our evaluation of the end-to-end embedded MT systems that we have available to us: transformation processes, document versions, and basic count measures. The transformation processes are listed vertically in the diamonds on the left side of figure 1. Starting with the hardcopy original document, each process transforms its input text and creates a new version. These document versions are listed vertically in the boxes in the second column of the figure. For each version, we compute one or more basic count measures on the words in that version's text. That is, for each process, there is an associated document version and for each document version, there are associated basic count measures. These count measures shown as A. through M. are defined in figure 2 below.

Two-Pass Evaluation

For each end-to-end system and language pair, we follow two separate passes in creating analysis files from scanned-in bitmap images. The first pass is for end-to-end Scan-OCR-MT evaluation: "OCR" the original document, then MT the resulting OCR-output file. The second pass is for Ground Truth-MT evaluation: "ground-truth" (GT) the original document, then MT the resulting GT-ed output file.

[4] We use "embedded MT systems" as defined in Voss and Reeder (1998).

[5] We chose a small scanner for portability of the system. Substituting in a flatbed scanner would not affect performance.

[6] See Voss and Van Ess-Dykema (2000) for a description of the products used.

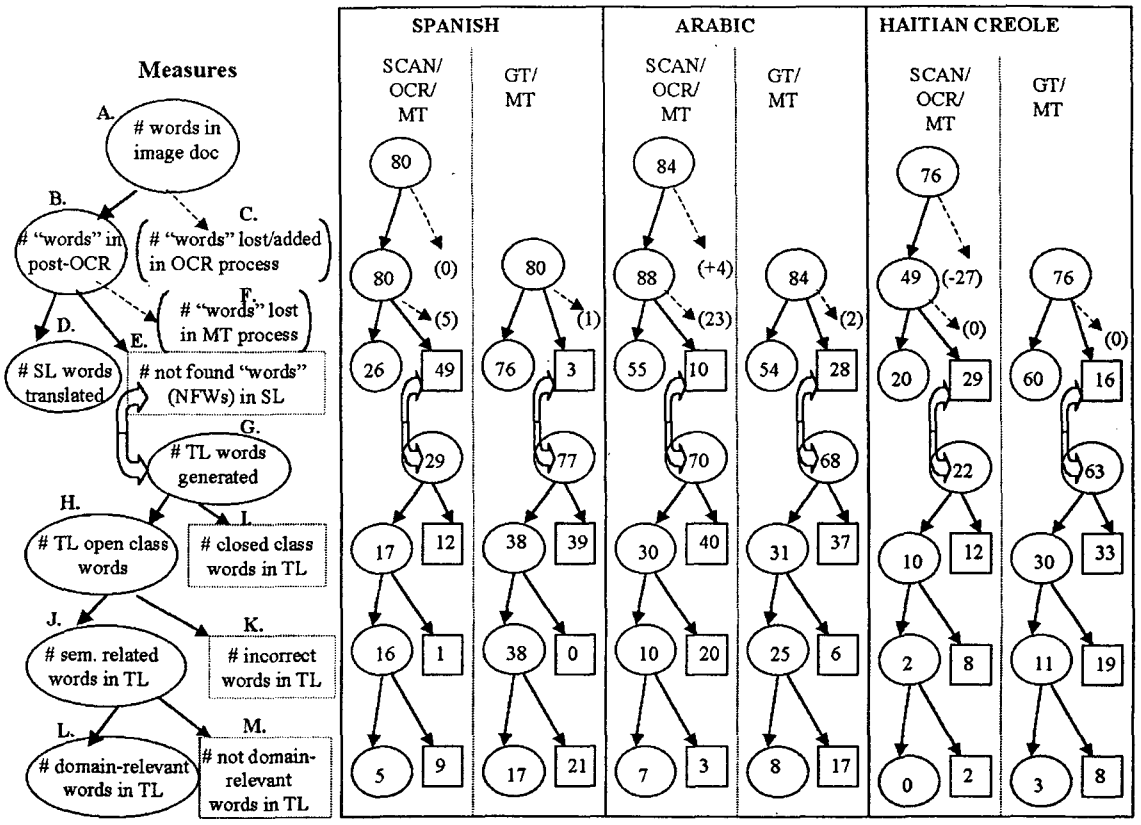

Figure 2 Comparison of Language-Specific System Results

The two passes represent the "worst" and "best" cases respectively for filtering within each of the three embedded MT systems. By "ground truth" versions of the document, we mean online duplicated versions that match, character-for-character, the input text.

We intentionally chose low-performance OCR software (for each language) to simulate a "worst case" performance by our systems, enabling us to compare them with the ideal high-performance ground-truth input to simulate a "best case" performance.

Texts from the Center for Disease Control

In order to compare the three language-specific systems, we had to find a corpus in a domain well-defined for filtering[7] that included parallel texts in Spanish, Arabic, and Haitian Creole. We found parallel corpora for these and many other languages at a website of the Center for Disease Control (CDC).[8]

We chose a paragraph from the chicken pox/varicella bulletin, page 2, for each of our three languages. This passage contains narrative full-length sentences and minimizes the OCR complications that arise with variable layouts. Our objective for selecting this input paragraph was to illustrate our methodology in a tractable way for multiple languages. Our next step will be to increase the amount of data analyzed for each language.

4 Analyses

We fill out one analysis template for each document tested in a language-specific system. Example templates with the basic count

[7] Filtering judgments are "well-defined" when multiple readers of a text in a domain agree on the "yes, relevant" status of the text.

[8] See http://www.immunize.org/vis/index.htm. The texts are "Vaccine Information Statements" describing basic medical symptoms that individuals should know about in advance of being vaccinated.

measures[9] are presented in figure 2 for each of the three embedded MT systems that we tested.

Notice that in figure 2 we distinguish valid words of a language from OCR-generated strings of characters that we identify as "words." The latter "words" may include any of the following: wordstrings with OCR-induced spelling changes (valid or invalid for the specific language), wordstrings duplicating misspellings in the source document, and words accurately OCR-ed. "Words" may also be lost in the MT process (see F.).[10]

The wide, block arrow in figure 2 connects E. and G. because they are both based on the MT output document. (We do not compute a sum for these counts because the E "words" are in the SL and the G words are in the TL.) The open class words (see H.) are nouns, verbs, adjectives, and adverbs. Closed class words (see I.) include all parts of speech not listed as open class categories.

In this methodology, we track the content words that ultimately contribute to the final filtering decision. Clearly for other tasks, such as summarization or information extraction, other measures may be more appropriate. The basic count measures A. through M. are preliminary and will require refinement as more data sets are tested. From these basic count measures, we define four derived percentage measures in section 5 and summarize these cases across our three systems in figure 3 of that section.

4.1 Embedded Spanish MT System Test

"Worst" case (Scan-OCR-MT pass)
As can be seen in figure 2, not all of the original 80 Spanish words in the source document retain their correct spelling after being OCR-ed. Only 26 OCR-ed "words" are found in the MT lexicon, i.e., recognized as valid Spanish words. Forty-nine of the OCR-ed "words" are treated as "not found words" (NFWs) by the MT engine, even though they may in fact be actual Spanish words. Five other OCR-ed "words" are lost in the MT process. Thus, the OCR process reduced the number of Spanish words that the MT engine could accept as input by more than 60%.

Of the remaining 40% that generated 29 English words, we found that 5 were "filter-relevant" as follows. The MT engine ignored 49 post-OCR Spanish "words" and working from the remaining 26 Spanish words, generated 29 English words.[11] Seventeen were open class words and 12 were closed class words. Nearly all of the open class words were translated correctly or were semantically appropriate for the domain (16 out of 17). From this correct set of 16 open class words, 5 were domain-relevant and 9 were not. That is, 5 of the 29 generated English words, or 17%, were semantically related and domain relevant words, i.e., triggers for filtering judgments.

"Best" case (GT-MT pass)
The MT engine generated 77 English words from the 80 original Spanish words. Thirty-eight, or half of the 77, were open class words; 39 were closed class words. All of the 38 open class words were correctly translated or semantically related to the preferred translation. And half of those, 17, were domain-relevant. Thus, the 77 English words generated by the MT engine contained 17 "filter-relevant" words, or 22%.

Comparing the Two Passes
Surprisingly the GT-MT pass only yields a 5% improvement in filtering judgments over the Scan-OCR-MT pass, even though the OCR itself reduced the number of Spanish words that the MT engine could accept as input by more than 60%. We must be cautious in interpreting the significance of this comparison, given the single, short paragraph used only for illustrating our methodology.

4.2 Embedded Arabic MT System Test

"Worst" case (Scan-OCR-MT pass)
The OCR process converted the original 84 Arabic words into 88 "words". Of the original 84 Arabic words in the source document, only

[9] The following formulas summarize the relations among the count measures: A = B+C; B = D+E+F; G = H+I; H = J+K; J = L+M.

[10] For example, we found that the word *la* in the Spanish text was not present in the TL output, i.e., the English equivalent *the* did not appear in the English translation.

[11] This occurred because the MT engine was not using a word-for-word scheme. The Spanish verb for *debo* is translated into 2 English words, *I must*. As we will note further on, different languages have different expansion rates into English.

55 retain their correct spelling after being OCR-ed and are found in the MT lexicon, i.e., recognized as valid Arabic words. Ten of the other OCR-ed "words" are treated as NFWs by the MT engine. The remaining 23 OCR-ed mixture of original words and OCR-induced "words" are not found in the Arabic MT lexicon. Thus, the OCR process reduced the number of original Arabic words that the MT engine could accept as input by slightly more than 65%.

Of the remaining 35% that generated 70 English words, we found that 7 were "filter-relevant" as follows. The MT lexicon did not contain 10 post-OCR Arabic "words" and working from the remaining 55 Arabic words, the MT engine generated 70 English words.[12] Thirty of the 70 were open class words and 40 were closed class words. Only one-third of the open class words were translated correctly or were semantically appropriate for the domain (10 out of 30). From this correct set of 10 open class words, 7 were domain-relevant and 3 were not. Thus, this pass yields 7 words for filtering judgments from the 70 generated English words, or 10%, were semantically related and domain relevant words.

"Best" case (GT-MT pass)
Of the 84 original Arabic words, even with the GT as input, 28 were not found in the MT lexicon, reflecting the engine's emerging status and the need for further development. Two others were not found in the Arabic MT lexicon, leaving 54 remaining words as input to the MT engine. The MT engine generated 68 English words from these 54 words. Thirty-one of the 68 were open class words; 37 were closed class words. Of the open class words, 25 were translated correctly or semantically related. And 8 of those 25 were domain-relevant. Thus, the 68 English words generated by the MT engine contained 8 "filter-relevant" words, or 12%.

Comparing the Two Passes
The GT-MT pass yields a 2% improvement in filtering judgments over the Scan-OCR-MT pass, even though the OCR itself reduced the number of Arabic words that the MT engine could accept as input by about 65%.

One of the interesting findings about OCR-ed Arabic "words" was the presence of "false positives," inaccurately OCR-ed source document words that were nonetheless valid in Arabic. That is, we found instances of valid Arabic words in the OCR output that appeared as different words in the original document.[13]

4.3 Embedded Haitian MT System Test

"Worst" case (Scan-OCR-MT pass)
In the template for the 76-word Haitian Creole source document, we see that 27 words were lost in the OCR process, leaving only 49 in the post-OCR document. Of those 49, only 20 exhibit their correct spelling after being OCR-ed and are found in the MT lexicon. Twenty-nine of the 49 OCR-ed "words" are not found (NFWs) by the MT engine. The OCR process reduced the number of original Haitian Creole words acceptable by the MT engine from 76 to 20, or 74%.

Of the remaining 26% that generated 22 English words, we found that none were "filter-relevant," i.e., 0%, as follows. The MT engine ignored 29 post-OCR "words" and working from the remaining 20 Haitian words, generated 22 English words. Ten were open class words and 12 were closed class words. Only 2 out of the 10 open class words were translated correctly or were semantically appropriate for the domain. From this correct set of 2 open class words, none were domain-relevant. The human would be unable to use this final document version to make his or her filtering relevance judgments.

"Best" case (GT-MT pass)
The MT engine generated 63 English words from the 76 original Haitian Creole words. Thirty of the 63 were open class words; 33 were closed class words. Only 11 of the 30 open class words were correctly translated or semantically related. Of those 11 words, 3 were domain-relevant. So, from the 63 generated English words, only 3 were "filter-relevant", or 5%.

[12] This expansion rate is consistent with the rule-of-thumb that Arabic linguists have for every one Arabic word yielding on average 1.3 words in English.

[13] As a result, the number of words in the two passes can differ. As we see in figure 2 in the Scan-OCR-MT pass, there were 55 SL words translated but, in the GT-MT pass, only 54 SL words in the original text.

Derived	Spanish		Arabic		Haitian Creole	
Meas.	OCR	GT	OCR	GT	OCR	GT
W.	40	95	35	64	26	79
X.	55	49	14	37	9	17
Y.	17	22	10	12	0	5
Z.	94	100	33	67	20	33

Figure 3 Summary of Language-Specific Results (percentages)

Comparing the Two Passes

With an OCR package not trained for this specific language and an MT engine from a research effort, the embedded MT system with these components does not assist the human on the filtering task. And even with the ground-truth input, the MT engine is not sufficiently robust to produce useful translations of valid Haitian Creole words.

5 Cross-System Results

In figure 3 we compare the three language-specific systems, we make use of four measures derived from the basic counts, A. through M., as defined in figure 2.

W. Original Document-MT Word Recall

% of original SL document words translatable by the MT engine after being OCR-ed. (D/A)

This measure on the GT pass in all 3 systems gives us the proportion of words in the original SL document that are in the individual MT lexicons. The Spanish lexicon is strong for the domain of our document (W = 95%). The measures for Arabic and Haitian Creole reflect the fact that their MT lexicons are still under development (W = 64% and 79%, respectively).

This measure on the OCR pass, given the corresponding measure on the GT pass as a baseline, captures the degradation introduced by the Scan-OCR processing of the document. From figure 3 we see that the Spanish system loses approximately 55% of its original document words going into the MT engine (95% minus 40%), the Haitian Creole 53% (79% minus 26%), and the Arabic 29% (64% minus 35%). Recall that the Spanish and Haitian Creole systems included the same OCR software, which may account for the similar level of performance here. This software was not available to us for Arabic.

X. MT Semantic Adequacy

% of TL words generated by MT engine that are open class & semantically adequate in their translation (J/G)

This measure is intended to assess whether a system can be used for filtering broad-level topics (in contrast to domains with specialized vocabulary that we discuss below). Here we see evidence for two patterns that recur in the two measures below. First, the GT pass---with one exception---exhibits better performance than the OCR pass. Second, there is a ranking of the systems with Haitian Creole at the low end, Arabic in the middle, and Spanish at the high end. We will need more data to determine the significance of the one exception (55% versus 49%).

Y. MT Domain-Relevant Adequacy

% of TL words generated by MT engine that are open class, semantically adequate in their translation, and domain-relevant (L/G)

In all of the systems there was a slight gain in domain-relevant filtering performance from the OCR pass to the GT pass. We can rank the systems with the Haitian Creole at the low end, the Arabic in the middle, and the Spanish at the high end: the measures in both the OCR and GT passes in Haitian Creole are lower than in the Arabic, which are lower than in the Spanish. Only the Spanish documents, but not the Arabic or Haitian Creole ones, when machine translated in either pass were judged domain-relevant by five people during an informal test.[14] Thus, our data suggests that the Spanish system's lower bound (OCR pass) of 17% on this measure is needed for filtering .

Z. MT Open Class Semantic Adequacy

% of open class TL words generated by MT engine that are semantically adequate in their translation (J/H)

[14] We are in the process of running an experiment to validate the protocol for establishing domain-relevant judgments as part of our research in measures of effectiveness (MOEs) for task-based evaluation.

The same pattern emerges with this measure. In each system there is an improvement in performance stepping from the OCR pass to the GT pass. Across systems we see the same ranking, with the OCR and GT passes of the Haitian Creole falling below the Arabic which falls below the Spanish.

Conclusion and Future Work

Our main contribution has been the proposal of an end-to-end process analysis template and a replicable evaluation methodology. We present measures to evaluate filtering performance and preliminary results on Spanish, Arabic and Haitian Creole FALCon systems.

The cross-system comparisons using the measures presented, with one exception, yielded the following expected rankings: (i) the GT-MT pass exhibits better performance than the Scan-OCR-MT pass and (ii) the Haitian Creole system is at the low end, Arabic is in the middle, and Spanish is at the high end.

Our long-term objective is to compare the results of the *system-internal* "measures of performance" (MOPs) presented here with results we still need from *system-external* "measures of effectiveness" (MOEs).[15] MOE-based methods evaluate (i) baseline unaided human performance, (ii) human performance using a new system and (iii) human expert performance. From this comparison we will be able to determine whether these two independently derived sets of measures are replicable and validate each other. So far, we have only addressed our original question, "when is an embedded MT system *good enough* for filtering?" in terms of MOPs. We found that, for our particular passage in the medical domain, documents need to reach at least 17% on our derived measure Y., MT domain-relevant adequacy (recall discussion of derived measure Y, in section 5).

Given that all but one process step ("ID wrong TL words" as shown in figure 1 where a human stick figure appears) in filling the template can be automated, the next phase of this work will be to create a software tool to speed up and systematize this process, improving our system evaluation by increasing the number of documents that can be regularly used to test each new system and reducing the burden on the operational linguists who assist us for the one critical step. Currently available tools for parallel text processing, including text alignment software, may provide new user interface options as well, improving the interactive assessment process and possibly extending the input set to include transcribed speech.

[15] See Roche and Watts (1991) for definitions of these terms.

Acknowledgements

We would like to acknowledge Lisa Decrozant (Army Research Laboratory) and Brian Branagan (Department of Defense) for language expertise and Francis Fisher (Army Research Laboratory) for systems engineering expertise.

References

Church, K. and Hovy, E. 1993. Good Applications for Crummy Machine Translation. *Machine Translation*, Volume 8, pages 239 - 258.

DARPA 1999. Translingual Information Detection, Extraction, and Summarization (TIDES) Initiative. http://www.darpa.mil/ito/research/tides/index.html

Fisher, F. and Voss, C. R. 1997. "FALCon, an MT System Support Tool for Non-linguists." In *Proceedings of the Advanced Information Processing and Analysis Conference*. McLean,VA.

Harmon, D. 1999. "A Framework for Evaluation in TIDES." Presentation at TIDES Planning Workshop, with link at http://www.dyncorp-is.com/darpa/meetings/tides99jul/agenda.html, July 28-30, Leesburg, VA.

Pomarede, J.-M., Taylor, K., and Van Ess-Dykema, C. 1998. Sparse Training Data and EBMT. In *Proceedings of the Workshop on Embedded MT Systems: Design, Construction, and Evaluation of Systems with an MT Component* held in conjunction with the Association for Machine Translation in the Americas (AMTA'98), Langhorne, PA, October.

Resnik, P. 1997. Evaluating Multilingual Gisting of Web Pages. In *Working Notes of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, Palo Alto, CA.

Roche, J. G. and Watts, B. D. 1991. Choosing Analytic Measures. *The Journal of Strategic Studies*, Volume 14, pages 165-209, June.

Sparck Jones, K. and Galliers, J. 1996. *Evaluating Natural Language Processing Systems*. Springer-Verlag Publishers, Berlin, Germany.

Taylor, K. and White, J. 1998. Predicting What MT is Good for: User Judgments and Task Performance. In *Proceedings of the Third*

Conference of the Association for Machine Translation in the Americas (AMTA'98), pages 364 -373, Langhorne, PA, October.

Voss, C. R. and Reeder, F. (eds.). 1998. *Proceedings of the Workshop on Embedded MT Systems: Design, Construction, and Evaluation of Systems with an MT Component* held in conjunction with the Association for Machine Translation in the Americas (AMTA'98), Langhorne, PA, October.

Voss, C. R. and Van Ess-Dykema, C. 2000. Evaluating Scan-OCR-MT Processing for the Filtering Task. Army Research Laboratory Technical Report, Adelphi, MD.