Automatic Domain Recognition for Machine Translation

Elke D. Lange & Jin Yang
SYSTRAN Software. Inc.
7855 Fay Avenue. Suite 300
La Jolla, CA 92037, USA
{elange, jyang}@systransoft.com

Abstract

This paper describes an ongoing project which has the goal of improving machine translation quality by increasing knowledge about the text to be translated. A basic piece of such knowledge is the domain or subject field of the text. When this is known, it is possible to improve meaning selection appropriate to that domain. Our current effort consists in automating both recognition of the text's domain and the assignment of domain-specific translations. Results of our implementation show that the approach of using terminology categorization already existing in the machine translation system is very promising.

1 Introduction

Along with the increasing growth of the World Wide Web, on-line translation has become an area of success for Machine Translation (MT) and also presents new challenges. It is agreed that, for a translation service catering to a wide audience, enhancement of robustness including automatic identification of language, domain and style level is required [2, 5, 9, 10]. The automatic setting of translation parameters has become one of our development priorities, thus increasing ease-of-use and translation quality at the same time. One area of great importance is the automatic selection of target language meanings depending on the domain of the text.

The automatic recognition of a text's domain is similar to text categorization, which can be defined as the context-based assignment of one or more predefined categories to texts [6]. Interest in automatic categorization of texts usually is related to the fields of information retrieval and machine learning. Papers on the subject of automatic domain recognition for MT were seldom found until recently when the problem became evident for the on-line translation environment. Another closely related research area is word sense disambiguation, which must be an MT developer's concern in providing high-quality translation. "This becomes very important when one is trying to further improve the quality of commercial MT systems which already have a store of knowledge required to produce fairly satisfactory translation"' [10]. We found two related works, which shared our motivation and goals of improving translation quality by recognizing domains. The "domain recognizer" in a Web-based English-Korean system [1] uses a text categorization technique, identifying 25 domains, and demonstrated a 45% accuracy for top domain and a 75% accuracy when the second top domain is applied. Another study [10] showed 12.0% improvement on the translation quality of Japanese-English MT.

The present paper describes an ongoing project that has the dual goal of automatically identifying the domain of a text and of improving MT quality by automatically linking domain identification to meaning selection. The goal of the project-Automatic Domain Recognition-is to increase translation accuracy by providing an automatic means of assigning domain-specific meanings to the translation of individual words. In order to do this, the system is first trained to recognize the domain of an entire or partial text, and on this basis, activate the appropriate domain-specific meanings for target language translation. Our current effort consists in automating both recognition of the text's domain and the assignment of domain-specific translations, and to base all recognition and meaning assignment algorithms on data already available in the system's lexical database.

2 Description

2.1 System Description

The SYSTRAN translation system, a general-purpose fully automatic MT system, employs a transfer approach. A unified and highly modular architecture applies to all language-pair systems [3, 4, 8], The dictionaries, an important integrated knowledge base for MT, not only contain bilingual lexicons for translation, but also other linguistic knowledge. Domain-

specific lexicons are part of SYSTRAN's source dictionaries. The domain differences are handled via terminology category, and the availability of a variety of domain-specific meanings. The information used for this project includes:

- SEMCAT: SEMantic CATegory. About 500 semantic categories are organized into six hierarchical trees. Lexical items may be assigned as many categories as necessary to define the semantic content of a word. The semantic codes are organized hierarchically to allow lower nodes to inherit the properties of all superior nodes. For example, English word *simulator* has code EQELE (electronic equipment), which is expanded to:

  EQELE → DEV → PHYSUB → INAM → PHENOM → THING

- TERMCAT: TERMinology CATegories. Represented as a flat structure, the TERMCAT set has 77 domain codes. Lexical items generally limited to a specific domain may be tagged with one of 77 domain codes. For example, the English words *astronomer, constellation, galaxy* and *meteoroid.* have the code ASTRONOMY.

- TG: Topical Glossary. Each entry in the system's dictionaries may be associated with several translations, each appropriate for another subject field. These domain-specific meanings are identified by a "Topical Glossary" (TG). Currently the system distinguishes approximately 30 Topical Glossaries, such as Aviation, Chemistry, Finance, Law etc. Users can manually set one or more TGs, in order of preference, when running a translation. For example, the following Chinese word has multiple English meanings coded in the dictionary.

  Chinese *di1ya1* "low pressure"

0	GENERAL	low pressure
1	PHYSICS	low pressure
2	ELECTRICITY	low tension
6	MEDICINE	minimum pressure
A	METEOROLOGY	low pressure

  Sometimes TG assignment is done even when the source language word is not polysemous. In the following example, the domain-specific TG translations indicate the probability of the translation in a specified domain as opposed to the translation in a general domain (e.g.. issue of currency or bonds in FINANCE; release of software in COMPUTER science).

  Chinese *falxing2* "distribute"

0	GENERAL	distribute
3	COMPUTER	release
e	FINANCE	issue

2.2 Translation Selection

For a transfer system, target language translation of lexical items is performed through a series of lexical transfer rules and bilingual dictionaries. In the lexicon (i.e., SYSTRAN's stem dictionary), a source language word has one general translation and optional domain-specific translations. While the general technical translation is the default, the domain-specific meanings can be activated via the TG selection parameter. For Chinese *di1ya1* "low pressure", the translation will be "minimum pressure" when TG=6 MEDICINE is selected. Target translation can also be assigned by word-specific linguistic rules (i.e.. SYSTRAN's expression dictionaries), in which there are extensive conditional lexical selection rules to assign translations based on the specified syntactic and/or semantic constraints. For example. Chinese *dilyal qi4liu2* "low-pressure air current" is coded in the expression dictionary as a collocation entry, thus *dilyal qi4liu2* "low-pressure air current" won't be translated as "minimum pressure air current" in the domain of MEDICINE.

Through the simple examples above, we can see that the assignment of target language translation is quite complex but hierarchical in the SYSTRAN systems. The hierarchy is motivated by manageability of large-scale systems [7]: the translation assignment in the stem dictionaries is simple and straightforward: the translation assignment in the expression level can be complex and sophisticated. For this paper, we limit our discussion to the former - single-word treatment.

3 Strategy

The problem of MT is one of accumulating and utilizing linguistic knowledge. Much of the needed knowledge is static and can be built into the system and its dictionary. Our strategy is based on the information in the dictionary. It is to utilize the dictionary information, including the rich semantic codes and terminology categories and to associate each domain with target language topical glossaries.

The strategy employed in automating domain recognition and translation assignment is as follows: an arbitrary input text is translated, the system's dictionaries and parsers associate SEMCATs and TERMCATs with individual words, a statistical analysis of the text determines the occurrence and frequency of these codes, and a "primary" and "secondary" domain are established on the basis of the most frequently identified domain codes. During the translation transfer phase a switch is set for the selection of TG meanings that are appropriate to the primary domain of the text, if none found in the dictionary, then the TG meaning associated with the secondary domain is taken.

This approach leverages on the already extensive information available in the system's dictionaries, but it also helps identify weaknesses of coverage in those dictionaries. This in turn provides an incentive to complete the dictionary coverage. Using some of the standard domain recognition techniques, namely ob-

taining high frequency words and phrases from domain-specific texts, we will feed information back into the dictionaries.

4 Testing

Four SYSTRAN language-pair systems were chosen for initial testing, i.e., Chinese-English, English-French, French-English and Russian-English. The systems are considered mature systems with production translation quality. They contain detailed linguistic rules and large terminology databases. The dictionaries for the four language-pair systems altogether contain over half a million terms.

A total of 650 files in the four source languages were downloaded from the Web. Thanks to the categorization provided by Yahoo! http://www.yahoo.com, Chinese. English and French texts in different domains are relatively easy to obtain. The data collection for Russian is relatively difficult. The following domains are used in the initial testing: AGRICULTURE. ASTRONOMY, AVIATION, BIOLOGY, CHEMISTRY, COMPUTER SCIENCE, DEFENSE, DRAMA, ECONOMY, FINANCE, HISTORY, LAW, LINGUISTICS, MATHEMATICS, MEDICINE, RELIGIONS, SPACE and SPORTS. The files vary, from popular science introduction (e.g., *What is civilization),* an organization's web page (e.g., *US Air Force News* http://www.af.mil/news/), to scientific articles (e.g., *The Molecular Anatomy of an Ancient Adaptive Event).* The average number of sentences of text is 125 sentences.

While searching and downloading files, the people were asked to select a primary domain for each file based on his/her judgement. The human-identified domain was considered the domain of the text, and was used to compare against the one automatically recognized. Next, we ran translations and evaluated the results based on both the recognition accuracy and the impact on translation.

5 Results

Since our current effort consists in automating both recognition of the text's domain and the assignment of domain-specific translations, we divide the evaluation process into accuracy of domain recognition and change in translation output.

5.1 Domain Recognition Accuracy

For each of the selected domains, there are 5 to 15 files. If one of the auto-detected domains is the same as the human-identified one, it is considered "correct" recognition. The accuracy for each domain is first measured:

Accuracy = Number of Correctly Recognized Files / Number of Files

Table 1 shows the accuracy of the domain recognition of the four language-pair systems. The left column (P) for each system shows the accuracy of primary domain recognition, and the right column (P+S) shows an accuracy when the human-identified domain was also identified by the system as either the primary or secondary domain. The overall accuracy is the average accuracy for each system, which is 77% (Chinese). 54% (English), 29% (French) and 48% (Russian).

	Chinese		English		French		Russian	
Domain	P	P+S	P	P+S	P	P+S	P	P+S
AGRICULTURE	80	100	67	83	100	100	100	100
ASTRONOMY	91	100	21	21	11	11	0	0
AVIATION	18	18	0	0	0	0	100	100
BIOLOGY	71	71	30	30	80	80	100	100
CHEMISTRY	80	90	21	36	0	75	-	-
COMPUTER	85	85	71	71	41	47	33	67
DEFENSE	7	7	0	0	0	0	-	-
DRAMA	46	64	0	0	0	0	100	100
ECONOMY	85	100	91	91	0	0	0	0
FINANCE	58	75	12	41	67	83	33	33
HISTORY	30	50	33	42	0	0	0	0
LAW	58	83	17	33	0	0	0	0
LINGUISTICS	70	100	75	92	-	-	-	-
MATH	-	-	44	61	0	0	0	0
MEDICINE	79	100	35	50	0	0	-	-
PEDAGOGY	-	-	-	-	-	-	100	100
RELIGION	100	100	8	17	57	71	25	50
SPACE	86	93	8	25	0	0	0	0
SPORTS	91	100	55	64	17	17	67	67
AVERAGE	77		54		29		48	

Table 1: Domain Recognition Accuracy (percentage)

The overview of the accuracy confirms some of our expectations.

- As expected, the accuracy of domain recognition relies on the dictionary information of each system. The more coverage of semantic information in a dictionary, the higher domain recognition accuracy will be. The rank of semantic code coverage of Chinese (90%), English (41%). French (34%) and Russian (50%) dictionaries resembles the overall domain recognition for the systems.

- Some domains are closely related to each other, such as DEFENSE and WARFARE. That Chinese has the low recognition accuracy (7%) in domain DEFENSE is due to the fact that the domains are mostly recognized as WARFARE. The connection among ECONOMY, TRADE and FINANCE, AVIATION, ASTRONOMY and SPACE etc. have a similar influence.

- The selection of domains is not a clear-cut task even for humans. The obvious example: texts about biochemistry, which go to either BIOLOGY or CHEMISTRY category. In this case, the texts were favorably recognized as BIOLOGY or CHEMISTRY as either the primary or secondary domains. The mystery often remains in HISTORY. Some files were recognized as WARFARE since the texts were about World War II or Napoleon Bonaparte; and one Chinese text was recognized as BOTANY since the text was about the history of tea in China.

5.2 Effect on Translation

The effect of recognizing the domain is visible in the translation. It comes from providing a substitute word that differs from the generalized word and thus fits the domain of the text. Figure 1 shows the impact on translation by calculating the percentage of sentences that show differences before and after activation of the domain recognition feature. The number of differences also reveals the coverage of domain-specific translation in the dictionaries of the four different systems

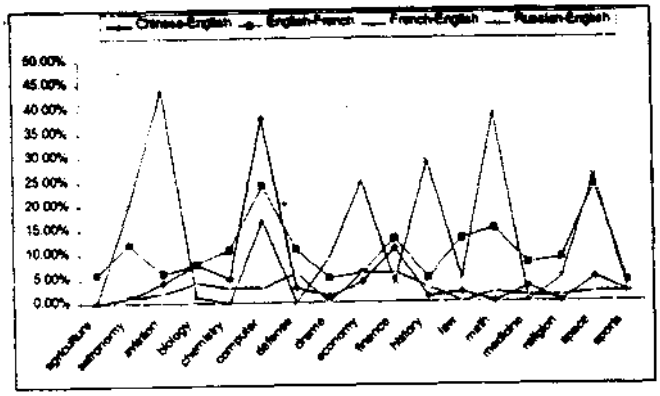

Figure 1: Translation Differences with Automatic Domain Recognition

The ranking according to translation differences found here, reflects a similar ranking in the evaluation of domain recognition. Both are in direct proportion to the extent of terminology category coverage in the respective dictionaries. Chinese, the newest of these systems, has the most extensive TERMCAT coverage, but still relatively poor TG coverage. The spike in Figure 1 for Chinese computer texts is the result of a concentrated and quick feedback from domain-specific texts to the system's dictionary (as mentioned in section 3. above).

Samples of translation differences are under review. The numbers from the Russian tests (Table 2) show how a substantial number of different word meanings were used in the translation due to the domain recognition. More importantly, the majority of changes are translation improvements. A more systematic analysis of translation is scheduled next.

File	P	S	Sent	Diff	Evaluation		
					better	worse	similar
hist	war	socio	80	25	+19	-8	~ 3
econ	peda	trade	56	10	+7	-1	~ 2
relig	relig	math	237	12	+7	-1	~ 5
avia1	avia	phy	46	17	+11	-5	~ 5
avia2	avia	war	112	52	+36	-4	~ 21
comsci	comsci	dyna	55	11	+6	-2	~ 3

Table 2: Translation Differences with Automatic Domain Recognition (Russian-English)

6 Conclusion

The preliminary results reported here are encouraging enough to show that automatic domain recognition based on a system's internal information is not only possible but also has beneficial translation results when coupled with meaning assignment techniques.

The current development also points out a need to boost the TERMCAT and TG coverage and identifies the particular areas of weakness in individual language systems. It also gave evidence of the benefit of such feedback in the example of Chinese COMPUTER terminology.

Several areas remain to be explored. We consider the following of primary importance:

- Expand the testing to more domains;
- Explore more expedient ways to boost TERMCAT coverage in the dictionaries;
- Experiment with domain recognition on the paragraph level:
- Move from single-word treatment to contiguous and non-contiguous expressions.

Recognizing the domain in which a text or portion of a text falls, is an important first step in the automatic MT process. Several levels of sophistication can then build on this in order to improve translation quality of single words and expressions.

The current development uses a simple approach to further utilize and enhance SYSTRAN's MT technology by improving the system's capability to select domain-specific translation of individual words. The results are both clear and promising. They have shown that translation accuracy increased in the translation of individual words. Moreover, they point to the possibility for domain specific rules and even the ability to consider the style of the text.

Acknowledgements

The work has been supported by NAIC (National Air Force Intelligence Center). We thank Dale Bostad of NAIC for his continuous interest and long-time support. We also thank Martha Awdziewicz and Uki MacIsaac of SYSTRAN Software, Inc. for their contribution.

References

[1] Sung-Kwon Choi, Han-Min Jung, Chul-Min Sim, Taewan Kim, Dong-In Park, Jun-Sil Park, and Key-Sun Choi. Hybrid Approaches to Improvement of Translation Quality in Web-based English-Korean Machine Translation. In *Proceedings of the 36th Annual Meeting of the*

Association for Computational Linguistics and 17th International Conference on Computational Linguistics, pages 251-255. Montreal. Quebec. Canada, 1998.

[2] Mary Flanagan. Two Years Online: Experiences, Challenges and Trends. In *Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas,* pages 206-211. Montreal, Quebec. Canada. 1996.

[3] Denis Gachot, Elke Lange, and Jin Yang. The SYSTRAN NLP browser: An application of machine translation technology in cross-language information retrieval. In Gregory Grefenstette. editor. *Cross-Language Information Retrieval,* pages 105-118. Kluwer Academic Publishers, 1998.

[4] Laurie Gerber and Jin Yang. SYSTRAN MT dictionary development. In *Machine Translation: Past, Present and Future: Proceedings of Machine Translation Summit VI,* pages 211-218, San Diego. CA. USA, 1997.

[5] Steve McLaughlin and Ulrike Schwall. Spicing Up the Information Soup: Machine Translation and the Internet. In *Machine Translation and Information Soup: Proceedings of Third Conference of the Association for Machine Translation in the Americas, AMTA '98,* pages 384-397. Langhorne, PA, USA, 1998.

[6] Isabelle Moulinier. A Framework for Comparing Text Categorization Approaches. In *AAAI Spring Symposium on Machine Learning and Information Access,* Stanford University, Stanford. CA. 1996.

[7] B. Scott. The Logos view. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas.* Columbia, MD, USA, 1994.

[8] Jin Yang and Laurie Gerber. SYSTRAN Chinese-English MT System. In *Proceedings of the International Conference on Chinese Computing '96.* Singapore. 1996.

[9] Jin Yang and Elke Lange. SYSTRAN on AltaVista. In *Machine Translation and Information Soup: Proceedings of Third Conference of the Association for Machine Translation in the Americas. AMTA '98,* pages 275-285, Langhorne. PA. USA, 1998.

[10] Yumiko Yoshimura, Kinoshita Satoshi, and Miwako Shimazu. Processing of Proper Nouns and Use of Estimated Subject Area for Web Page Translation. In *Proceedings of The 7th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-97),* pages 10-18, Santa Fe, New Mexico, USA, 1997.