Improving Statistical Machine Translation in the Medical Domain using the Unified Medical Language System

Matthias Eck **Stephan Vogel** **Alex Waibel**

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213, USA

matteck@cs.cmu.edu vogel+@cs.cmu.edu ahw@cs.cmu.edu

Abstract

Texts from the medical domain are an important task for natural language processing. This paper investigates the usefulness of a large medical database (the Unified Medical Language System) for the translation of dialogues between doctors and patients using a statistical machine translation system. We are able to show that the extraction of a large dictionary and the usage of semantic type information to generalize the training data significantly improves the translation performance.

1   Introduction

Hospitals in the United States have to deal with an increasing number of patients who have no knowledge of the English language. It is not surprising that in this area translation errors can lead to severe problems (Neergard, 2003; Flores et al. 2003). This is one of the main reasons why the medical domain plays an important role in many of the current projects involving natural language processing. Especially many text or speech translation projects include tasks to translate texts or dialogues with medical topics.

The goal of this research was the improvement of translation quality in the medical domain using a statistical machine translation system. A statistical machine translation system deduces translation rules from large amounts of parallel texts in the source and target language.

The general approach to gather as much training data as possible is usually complicated and expensive. So it is necessary to make use of already available data and databases and it is reasonable to hope that some ideas and special methods could actually improve the performance in limited domains, like the medical domain.

The Internet and especially the WWW offers a lot of data related to medical topics. Especially interesting and promising for us was the Unified Medical Language System® (UMLS, 1986-2004) available from the US National Library of Medicine. It provides a vast amount of information concerning medical terms and we extracted information from this database to improve an existent translation system.

The paper will first give an introduction into the Unified Medical Language system. We will then point out which parts could be useful for statistical machine translation and later show how the baseline system was actually significantly improved using this data.

2   The Unified Medical Language System

2.1   Introduction

The Unified Medical Language System (UMLS, 1986-2004) project was initiated in 1986 by the U.S. National Library of Medicine. It integrates different knowledge sources into one database (e.g. biomedical vocabularies, dictionaries).

The goal is to help health professionals and researchers to use biomedical information from these different sources. It is usually updated about 3 or 4 times per year.

It consists of three main knowledge repositories, the UMLS Metathesaurus, the UMLS Semantic Network and the SPECIALIST lexicon.

Interesting facts about the UMLS, related work and further information can be found in (Lindbergh, 1990; Kashyap, 2003; Brown et al., 2003; Friedman et al., 2001; Zweigenbaum et al., 2003).

2.2   The UMLS Metathesaurus

The UMLS Metathesaurus provides a common structure for approximately 100 source biomedical vocabularies.

The 2003AB[1] version of the Metathesaurus contains exactly 900,551 concepts named by 2,247,457 terms. It is organized by concept, which is a cluster of terms (i.e. synonyms, lexical variants

---

[1] 2003AB was the actual release when the experiments described in this paper were executed. The most recent version now is 2004AA, which contains certain additional and updated information. All numbers given in this paper are according to the 2003AB version.

and translations) with the same meaning. Translations are present for up to 14 additional languages besides English. It is very likely that other languages will be added in later releases.

Table 1 shows the distribution of the terms according to the 15 different languages.

Language	Number of Terms
English	1860683
Spanish	73136
German	71316
Portuguese	69127
Russian	44907
Dutch	38600
French	38249
Italian	24992
Finnish	22382
Danish	723
Swedish	723
Norwegian	722
Hungarian	718
Basque	695
Hebrew	484

Table 1: Languages in the UMLS

For example the concept "arm" includes the English lexical variant, its plural form, "arms" and with "bras", "arm", "braccio", "braco", "ruka" and "brazo" the French, German, Italian, Portuguese, Russian and Spanish translations.

Some entries contain case information, too, and the entries are not limited to words but some terms are also longer phrases like "third degree burn of lower leg" or "loss of consciousness".

It also includes inter-concept relationships across the multiple vocabularies. The main relationship types are shown in Table 2:

Relationship types
broader
narrower
other related
like
parent
child
sibling
is allowed qualifier
can be qualified by
is co-occurring with

Table 2: Relationship types

The synonym-relationship is implicitly realized by different terms that are affiliated with the same concept.

The co-occurrence relationship refers to concepts co-occurring in the MEDLINE-publications.

In addition each concept is categorized into semantic types according to the UMLS Semantic Network.

2.3 The UMLS Semantic Network

The UMLS Semantic Network categorizes the concepts of the UMLS Metathesaurus through semantic types and relationships.

Every concept in the Metathesaurus is part of one or more semantic types.

There are 135 semantic types arranged in a generalization hierarchy with the two roots "Entity" and "Event". This hierarchy is still rather abstract (e.g. not deeper than six).

A more detailed generalization hierarchy is realized with the child, parent and sibling relationships of the UMLS Metathesaurus.
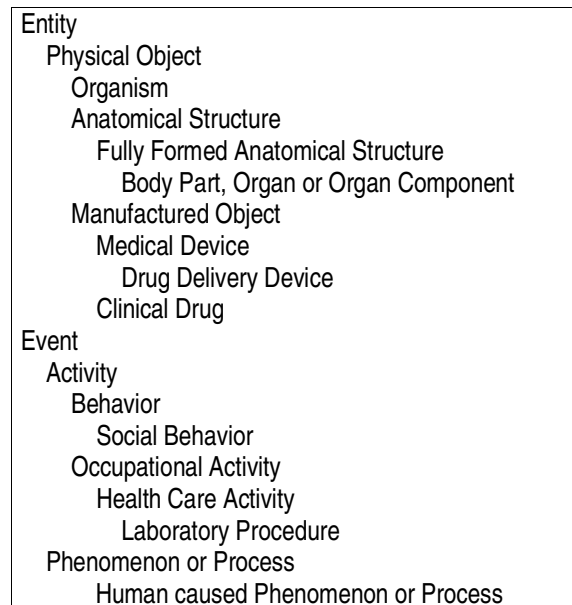
Figure 1 shows some examples for semantic types.

```
Entity
  Physical Object
    Organism
    Anatomical Structure
      Fully Formed Anatomical Structure
        Body Part, Organ or Organ Component
    Manufactured Object
      Medical Device
        Drug Delivery Device
      Clinical Drug
Event
  Activity
    Behavior
      Social Behavior
    Occupational Activity
      Health Care Activity
        Laboratory Procedure
  Phenomenon or Process
      Human caused Phenomenon or Process
```

Figure 1: Some semantic types

2.4 The SPECIALIST lexicon

The SPECIALIST lexicon contains over 30,000 English words. It is intended to be a general English lexicon including many biomedical terms.

The lexicon entry for each word or term records the syntactic, morphological and orthographic information.

```
{base=anesthetic
spelling_variant=anaesthetic
entry=E0330018
    cat=noun
    variants=reg
    variants=uncount
}
```

Figure 2: Example entry from the Specialist Lexicon

Figure 2 shows the entry for "anesthetic". There is a spelling variant "anaesthetic" and an entry number. The category in this case is noun (there is another entry for "anesthetic" as an adjective). The variants-slot contains a code indicating the inflectional morphology of the entry. "anesthetic" can either be a regular count noun (with regular plural "anesthetics") or an uncountable noun.

3 Machine Translation Experiments

3.1 The Baseline System

The Baseline system, which we used to test different approaches to improve the translation performance, is a statistical machine translation system. The task was to facilitate doctor-patient dialogues across languages. In this case we chose translation from Spanish to English.

The Baseline system was trained using 9,227 lines of training data (90,012 English words, 89,432 Spanish words). 3,227 lines of this data are "in-domain" data. We collected doctor patient dialogues during ongoing research projects in our group and used this data as training data. The 6,000 other lines of training data are out of domain data from the C-Star Project. This data also consists of dialogues but not from the medical domain.

The test data consists of 500 lines with 6,886 words. The test data was also taken from medical dialogues between a doctor and a patient and contains a reasonable number of medical terms but the language is not very complex. Figure 3 shows some example test sentences (from the reference data).

	(…)
Doctor:	The symptoms you are describing and given your recent change in diet, I believe you may be anemic.
Patient:	Anemic? Really? Is that serious?
Doctor:	Anemia can be very serious if left untreated. Being anemic means your body lacks a sufficient amount of red blood cells to carry oxygen through your body.
	(…)

Figure 3: Example test sentences (reference)

The Baseline system uses IBM1 lexicon transducers and different types of phrase transducers (Zhang et al. 2003, Vogel et al. 1996, Vogel et al. 2003). The Language model is a trigram language model with Good-Turing-Smoothing built with the SRI-Toolkit (SRI, 1995-2004) using only the English part of the training data.

The Baseline system scores a 0.171 BLEU and 4.72 NIST. [BLEU and NIST are well known scoring methods for measuring machine translation quality. Both calculate the precision of a translation by comparing it to a reference translation and incorporating a length penalty (Doddington, 2001; Papineni et al., 2002).]

3.2 Extracting dictionaries from the UMLS

The first way to exploit the UMLS database for a statistical machine translation system naturally is to extract additional Spanish-English lexicons or phrasebooks.

The UMLS Metathesaurus provides translation information as we can assume that Spanish and English terms that are associated with the same concept are respective translations. For example as the English term "arm" is associated with the same concept as the Spanish term "brazo" we can deduce that "arm" is the English translation of "brazo".

Unfortunately the UMLS does not contain morphological information about languages other than English. This means it cannot be automatically detected that "brazo" is the singular form and thus the translation of "arm" and not the translation of "arms".

As most of the entries are in singular form we just extracted every possible combination of Spanish and English terms regardless of possible errors like combining the singular "brazo" and the plural "arms".

The resulting (lower-cased) Spanish-English lexicon/phrasebook contains 495,248 pairs of

words and phrases. This means each Spanish term is combined with seven English terms on average.

This seems to be an extremely huge amount but it has to be considered that there are terms in the UMLS and the resulting lexicon that are probably too special to be really useful for the translation of dialogues (e.g. "1,1,1-trichloropropene-2,3-oxide" translating to "oxido de tricloropropeno").

Nevertheless there are lots of meaningful entries as the following experiments show.

Applying the dictionaries to the Baseline system
In the first step we just added this lexicon/phrasebook as an additional transducer and did not change the language model.

The experiment showed a nice increase in BLEU and NIST performances and scored at 0.180 BLEU and 4.86 NIST.

This system especially has a higher coverage, as only 302 words (types) are not covered by the training data compared to 411 for the baseline system.

Adding the English side to the Language Model
As the extracted dictionary contained many phrases it seemed reasonable to add the English side to the language modeling data. This also prevents words from the extracted dictionary to be treated as "unknown" by the language model if they were not in the language model training data. This further improved the BLEU and NIST scores to 0.182 BLEU and 4.92 NIST.

It should not be surprising to get an improvement in these first two experiments because basically just more data was used to train the systems. The really interesting ideas will be presented in the next sections.

3.3 Using the Semantic Type Information

The overall idea to use the semantic type information is to generalize the training data.
The training data contains for example sentence pairs like:

Necesito examinar su *cabeza.*
I need to examine your *head.*
Necesito examinar su *brazo.*
I need to examine your *arm.*
Necesito examinar su *rodilla.*
I need to examine your *knee.*

If we could generalize these sentences by replacing the special body parts like "head", "arm" and "knee" with a general tag e.g. "@BODYPART" and especially treat this tag we could use one sentence of training data for every body part imaginable in this sentence.
We would just need an additional lexicon that just translates body parts.

Necesito examinar su @BODYPART.
I need to examine your @BODYPART.

We could additionally correctly translate possibly unseen sentences like "Necesito examinar su *antebrazo*" ("I need to examine your *forearm*") if we could automatically deduce that "antebrazo/forearm" is a body part and if we just knew this translation pair.

Some additional similar sentences in which we could apply the same ideas are:

Enseneme que @BODYPART es.
Show me which @BODYPART.
¿Que @BODYPART le/la duele?
Which @BODYPART hurts?

(In the last sentence it actually depends on the gender of the body part on the Spanish side if the sentence is "¿Que @BODYPART *la* duele?" or "¿Que @BODYPART *le* duele?". But as we are translating from Spanish to English this did not seem to be a big problem.)

As stated before every concept in the UMLS Metathesaurus is categorized into one or more semantic types defined in the UMLS Semantic Network.

The two semantic types "Body Part, Organ, or Organ Component" and "Body Location or Region" from the UMLS Semantic Network cover pretty closely what we usually affiliate with the colloquial meaning of *body part*.

[The terminological difference is that the semantic type "Body Part, Organ, or Organ Component" is defined by a certain function. For example "liver" and "eye" are part of this semantic type, whereas the semantic type "Body Location or Region" is defined by the topographical location of the respective body part. Examples are "head" and "arm". The function in this case is not as clearly defined as the function of a "liver".]

This information was used in the next experiment. We first filtered the general Spanish-English dictionary, we had extracted from the UMLS, to contain only words and phrases from the two semantic types "Body Part, Organ, or Organ Component" and "Body Location or Region". This gave a dictionary of 11,260 translation entries for

body parts. Again each Spanish term is combined with about seven English terms on average.

In the next step we replaced every occurrence of a word or phrase pair from this new dictionary in the training data (i.e. if it occurred on the Spanish and English side) with a general body-part-tag.

527 sentence pairs of the original 9,227 sentence pairs contained a word or phrase pair from this dictionary.

A retraining of the translation system with this changed training data resulted in transducer rules containing this body-part-tag.

By using cascaded transducers (Vogel and Ney, 2000) in the actual translation the first transducer, that is applied (in this case the body-part dictionary) replaces the Spanish body part with its translation pair and the body-part tag.

The following transducers can apply their generalized rules containing the body-part-tag instead of the real body part.

E.g. translation of the sentence:

Necesito examinar su antebrazo.

First step apply body-part dictionary rule (antebrazo→forearm)

Necesito examinar su @BODYPART(antebrazo→forearm).

Apply generalized transducer rule: (a rule could be: Necesito examinar su @BODYPART → I need to examine your @BODYPART)

I need to examine your @BODYPART(antebrazo→forearm).

Resolve tags:

I need to examine your forearm.

By applying this to the whole translation system the score improved to 0.188 BLEU/4.94 NIST.

Using other semantic types

As the body-part lexicon and the replacement of body-parts proved to be helpful we applied two more of these replacement strategies. Consider the following 4 sentence pairs from the training data.

| ¿Siente *dolor* cuando respira? |
| Do you feel *pain* when you breathe? |
| ¿Cuando le empezo la *fiebre*? |
| When did the *fever* start? |
| ¿Podria ser *artritis*? |
| Could this be *arthritis*? |
| ¿Es grave la *anemia*, doctor? |
| Is *anemia* serious, doctor? |

The first two sentences contain findings or symptoms with the terms "dolor/pain" and "fiebre/fever". The second two sentences contain diseases with "artritis/arthritis" and "anemia/anemia". The appropriate semantic types from the UMLS Semantic Network for these terms are "Finding" and "Sign or Symptom" for "pain" and "fever" and "Disease or Syndrome" for "arthritis" and "anemia"

Filtering the Spanish-English dictionary resulted in 25,987 "Finding/Sign or Symptom" translation pairs (approximately three English terms per Spanish term) and 116,793 "Disease or Syndrome" translation pairs (approximately five English terms per Spanish term).

198 sentence pairs from the training data contained a "Finding/Sign or Symptom"-pair and 127 sentence pairs contained a "Disease or Syndrome"-pair from these dictionaries.

The final translation with those three semantic types replaced in the training data and using the three filtered dictionaries with the cascaded transducer application gave a translation performance of 0.190 BLEU/5.02 NIST.

This shows that although less than 10% of the sentences were affected by the replacement with the appropriate tags we could nicely improve the overall translation performance.

Example translations

Some example translations comparing the baseline and the best system with the reference are listed in table 3.

1. Sentence Reference	the condition is called tenosynovitis, which is an inflammation of the tendon sheath.
Baseline	this condición diagnostic, which is a inflammation from the of the tendon.
Best System	this condition is called tenosynovitis, which is a inflammation of tendon sheath.
2. Sentence Reference	i guess your work involves a lot of repetitive movement, huh?
Baseline	do you I guess your work require plenty baby´s, no?
Best System	i guess you your work require plenty repetitive movements, not?
3. Sentence Reference	you need vitamin c and iron in your blood to help your body
Baseline	you need vitamin c and iron in your blood help rescue to
Best System	you need vitamin c and iron in your blood to help the body
4. Sentence Reference	did you take anything for the pain?
Baseline	did you sleep taken anything for the pain?
Best System	did you taken anything for the pain?
5. Sentence Reference	i can feel it here, behind my breastbone.
Baseline	i here, behind of the esternón.
Best System	i here, behind of sternum.

Table 3: Example translations

The last example sentence is an interesting case. The best system does not get more words right compared to the baseline system and so the BLEU/NIST-score does not improve. But "sternum" is a synonym of the correct "breastbone" and a more technical term. This supports the claim that the UMLS tends to contain more technical terms (like "tenosynovitis" in the first sentence).

4 Future work

It is surely possible to use every semantic type from the semantic network in the same way like the overall five semantic types, which were used in the experiments. We did not do this here because further semantic types occurred extremely rarely in the test and training data. But this could easily be done for other test and training data and it is reasonable to expect similar improvements.

Another idea is to use a more specialized approach and to make use of the relationships in the UMLS Metathesaurus. Each concept could be generalized by its parent-concepts instead of its semantic type. The generalization hierarchy for the concept "leg" is for example: leg → lower extremity → extremity → body region → anatomy.

This could be especially helpful when translating to morphologically richer languages than English because the usage of extremities could differ from other body parts for example.

In the extracted dictionaries every translation pair was given the same translation probability. It might be helpful to re-score these probabilities by using information from bilingual or monolingual texts to improve the translation probabilities for usually frequently used terms compared to rarely used terms.

As the example translations showed, the extracted dictionaries from the UMLS tend to contain technical terms instead of colloquial terms (translation "sternum" instead of "breastbone"). We can further assume that a doctor prefers to use the more technical terms and a patient prefers the more colloquial terms. Therefore it could be interesting to examine if having two different translation systems for sentences uttered by a doctor and a patient would improve the overall translation performance.

5 Conclusion

We carried out four different experiments in order to improve a Spanish-English medical domain translation system. After sequentially applying different ideas the final system shows an 11% improvement in BLEU and 6% improvement in NIST score.

Table 4 compares the different experiments and scores (the 500k dictionary refers to the dictionary that was first extracted from the UMLS with 495,248 word pairs).

System	BLEU	NIST
Baseline system	0.171	4.72
+500k dictionary	0.180	4.86
+LM improvement	0.182	4.92
+body part-tags	0.188	4.94
+sign/symptom/finding +disease/syndrome	0.190	5.02

Table 4: Experiments and improvements

With more investigation and the ongoing effort of the National Library of Medicine to extent the UMLS databases it will hopefully be possible to further improve the translation performance.

References

Allen C. Browne, Guy Divita, Alan R. Aronson, Alexa T. McGray, 2003. *UMLS Language and Vocabulary Tools*, Proceedings of the American Medical Informatics Association (AMIA) 2003 Symposium, Washington, DC, USA.

George Doddington. 2001. *Automatic Evaluation of Machine Translation Quality using n-Gram Cooccurrence Statistics*. NIST Washington, DC, USA.

Glenn Flores, M. Barton Laws, Sandra J. Mayo, Barry Zuckerman, Milagros Abreu, Leonardo Medina,
Eric J. Hardt, 2003. *Errors in medical interpretation and their potential clinical consequences in pediatric encounters*, Pediatrics, Jan 2003.

Carol Friedman, Hongfang Liu, Lyuda Shagina, Stephen Johnson, George Hripcsak, 2001. *Evaluating the UMLS as a Source of Lexical Knowledge for Medical Language Processing*, Proceedings of the AMIA 2001 Symposium, Washington, DC, USA.

Vipul Kashyap, 2003. *The UMLS semantic network and the semantic web*, Proceedings of the AMIA 2003 Symposium, Washington, DC, USA.

C. Lindberg, 1990. *The Unified Medical Language System (UMLS) of the National Library of Medicine*, Journal of the American Medical Record Association, 1990;61(5):40-42.

Lauren Neergard, 2003. *Hospitals struggle with growing language barrier*, Associated Press, The Charlotte Observer Sept. 2, 2003

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation,* Proceedings of the ACL 2002, Philadelphia, USA.

SRI Speech Technology and Research Laboratory, SRI Language Modeling Toolkit, 1995-2004 (ongoing)
http://www.speech.sri.com/projects/srilm/

UMLS Unified Medical Language System, National Library of Medicine, 1986-2004 (ongoing)
http://www.nlm.nih.gov/research/umls/

Stephan Vogel and Hermann Ney, 2000. *Translation with Cascaded Finite State Transducers*. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000), pp. 23-30. Hongkong, China, October 2000.

Stephan Vogel, Hermann Ney, and Christoph Tillmann, 1996. *HMM-based Word Alignment in Statistical Translation*, Proceedings of COLING 1996: The 16th International Conference on Computational Linguistics, pp. 836-841. Copenhagen, August 1996.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venogupal, Bing Zhao, Alex Waibel, 2003. *The CMU Statistical Translation System*, Proceedings of MT-Summit IX. New Orleans, LA. Sep 2003.

Ying Zhang, Stephan Vogel, Alex Waibel, 2003. *Integrated Phrase Segmentation and Alignment Algorithm for Statistical Machine Translation*, Proceedings of International Conference on Natural Language Processing and Knowledge Engineering 2003, Beijing, China, Oct 2003.

Pierre Zweigenbaum, Robert Baud, Anita Burgun, Fiammetta Namer, Éric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Benoît Thirion, Stéfan Darmoni, 2003. *UMLF: a Unified Medical Lexicon for French*, Proceedings of the AMIA 2003 Symposium, Washington, DC, USA.