Conceptual transfer: Using local classifiers for transfer selection

Gregor Thurmair
gregor.thurmair@gmx.de

Abstract

A key challenge for Machine Translation is transfer selection, i.e. to find the right translation for a given word from a set of alternatives (1:n). This problem becomes the more important the larger the dictionary is, as the number of alternatives increases. The contribution presents a novel approach for transfer selection, called *conceptual transfer*, where selection is done using classifiers based on the conceptual context of a translation candidate on the source language side. Such classifiers are built automatically by parallel corpus analysis: Creating subcorpora for each translation of a 1:n package, and identifying correlating concepts in these subcorpora as features of the classifier. The resulting resource can easily be linked to transfer components of MT systems as it does not depend on internal analysis structures. Tests show that conceptual transfer outperforms the selection techniques currently used in operational MT systems.

Keywords: Machine translation, transfer selection, rule-based MT

1. Rationale[1]

Corpus-based techniques for lexicon creation in monolingual and bilingual settings are widely used nowadays to improve lexicon coverage and adequacy (cf. Thurmair & Aleksić 2012). If such lexica are merged with already existing ones, two effects can be observed: Creation of completely new entries (a new source with its translation(s)), and (much more often) creation of new translations for an already known source entry. The later effect increases the number of translation alternatives for a given source word. If the underlying MT system is not capable to select the right translation in a given context from this larger set of alternatives, then the lexicon increase has no effect; in some cases the results can even deteriorate. Therefore, the issue of transfer selection becomes the more pressing the larger the lexicons are.

1.1 Current techniques for transfer selection

Transfer selection in current MT uses one of the following techniques:

1. **Lexicon Module** sequence: The lexicon is divided into different modules, and the users define the sequence of searching in the translation settings (e.g.: '*user lexicon'* before '*company lexicon'* before '*system lexicon'*).

This strategy has limited scope if translation alternatives are stored in the *same* module. Also, it does not reflect the contextual situation in which a transfer selection is required.

2. **Global variable settings**: Global variables like domain information (like '*engineering'*, '*medical'*, '*legal'*, '*IT'* domain) are set for a text (by the users or by automatic topic identification), and the lexicon is made sensitive to the domain selection by tagging the single translations with such domain tags, which then are preferred. For instance, (de) '*Läufer'* -> (en) '*runner'* but -> ['*chess'*] '*bishop'*.

However, domain information as used in terminology is of only limited use for MT: In 1:1 cases ((de) '*Lumbalanästhesie*' -> (en) '*intraspinal anesthesia*') the domain tag ['*MED*'] is superfluous, while in other cases (1:m) it is not sufficient: (de) '*Anlage*' -> (en) '*investment*' ['*Finance*']; however even in finance texts there are papers with an (de) '*Anlage*', to be translated as (en) '*appendix*'. So, domain information fails in cases where different readings must be activated in the same domain and local contexts.

Tests in PANACEA with automotive texts (Aleksić et al. 2013) have shown that using domain tags can even deteriorate quality because of overspecification, i.e. cases where the automotive-specific term is used in a more general context. Examples are:

(de) '*Rahmen*' -> (en) [*AUTO*] '*chassis*' but (de) '*Rahmensystem*' -> '*frame system*' (*'*chassis system*')

(de) '*Leitung*' -> (en) [*AUTO*] '*pipe*' but '*Leitung der Firma*' -> '*company management*' (*'*company pipe*')

(de) '*Mangel*' -> (en) [*AUTO*] '*fault*' but '*Ingenieur-mangel*' -> '*lack of engineers*' (*'*engineer fault*')

Such mistakes balanced out the terminology improvements, and the overall translation quality did not increase significantly.

3. A third heuristic is to provide **contextual tests** (Thurmair 1990). Such tests refer to morphological, syntactic, semantic, discourse and other contexts; e.g.: (en) '*eat*': [if subject is human] -> (de) '*essen*'; [if subject is animated] -> (de) '*fressen*'). They can be rather complex (Eberle 2008). Such tests are specified in the lexicon as transfer selection conditions. At runtime, the tests compare the test conditions of the lexicon with an underlying representation of the input sentence. The respective transfer entry is selected if the test succeeds, i.e. if it matches this underlying tree.

This kind of tests has two drawbacks: (a) It has only limited coverage: not all transfer distinctions can be expressed by such tests. (b) But the main problem with this heuristics is robustness: Tests fail if the underlying syntactic structure built by the system is incorrect (parse errors). Then the right translation is not returned.

[1] This research has been supported by the EU research program, project PANACEA, FP7-ICT-2009-4-248064. The resource is available in METASHARE, http://www.meta-net.eu/meta-share

4. Other approaches have been tried based on simple **corpus frequency**; but they fail when different translations need to be used depending on the sentential context.
5. In statistical MT, context is taken into account in the decoding. However, the transfer is domain (and even training-corpus) dependent, new domains must be trained anew (Pecina et al. 2012). Also, transfer selection is difficult to influence by users (Itagaki et al. 2007). Moreover, only a fraction of the translations offered in a standard dictionary can be covered by such methods, as will be shown below.
As a result, the current transfer selection strategies are not able to cope with the requirements to use large bilingual lexicons as can be produced nowadays from aligned corpora. But as long as there is no way to select the best transfers for a given context from a number of options, there is no point in enlarging the translation dictionaries with material which cannot be used by the system.

2 Conceptual transfer

Transfer in MT can be divided into *structural* transfer (independent of lexical material, like: fronting of adjectives in English) and *lexical* transfer. Lexical transfer can be *simple* (word-by-word replacement) or *complex*, if it depends on the context. Such context can be syntactic (like the semantic marker of the subject node in the '*eat*' -> '*essen*' vs. '*fressen*' example above); but they can also be taken just from the concepts which surround a given translation candidate.
Conceptual transfer is a method to select the translation of a candidate on the basis of the (source language) concepts with which it co-occurs. The context is local, i.e. just the sentence of paragraph surrounding the candidate is used (as opposed to global domain settings).

2.1 Approach

The approach taken here tries to model human intuition, which is able to determine how a term should be translated by looking at its conceptual context (i.e. the words surrounding it). The idea of the conceptual transfer is to identify such conceptual contexts, using parallel corpora.
The task starts from two resources: 1. a *lexicon* containing possible transfers of a given word; 2. a *parallel corpus* which allows to identify context concepts for its different translations.
The task is executed in the following way:
1. Take a bilingual dictionary, and identify the packages they contain (‚package' being a set of translations sharing the same source lemma and POS); such packages are the target of the disambiguation effort.
2. For each target lemma in each package, create a subcorpus of sentences containing this lemma as the translation.
3. Build a classifier for each of these translations, based on the (source language part of the) respective subcorpora. The features of the classifier are the best co-occurring source language terms for a given translation candidate.
At runtime, this classifier will be used to find the best possible translation of a word in the given local context.

2.2 Related work

1. There is significant research in **learning transfer rules**. (Caseli et al., 2008, Sánchez-Martínez et al. 2007, 2009a; Tyers et al. 2012), (Probst et al., 2002; Probst, 2005, Lavie et al., 2008; Hannemann et al., 2009, Menezes & Richardson (2001). Unlike the present approach, this research focuses on the aspects of structural (lexicon independent) transfer, and less on the aspect of lexical selection[2].
2. Insights can also be found in **word sense** disambiguation. Here, for the definition of a word sense inventory, it has been proposed to use multilingual material. (Resnik and Yarowsky, 1997[3], Ide et al. (2002), Miháltz (2005) , Apidianaki (2008)). The approach assumed that word senses correlate to different translations, but it has been shown (Specia et al., 2006) that this relation does not really hold: Polysemous words like (de) '*Zelle'* (in biology, energy, politics, cloister) transfer *all* their meanings into one single target (en) '*cell';* in turn, the same meaning (de) '*ausschlafen*' translates into two concepts (en) '*sleep in*' and '*sleep out*'. The current approach is similar to this research in the attempt to use conceptual contexts for analysis. However it does not intend to disambiguate word senses but to find the best transfer for a given word from a set of candidates[4].
3. An approach of disambiguation of source language contexts was presented in Thurmair (2006), called 'neural transfer' there. A *monolingual* corpus was used, the translation candidates were *manually* annotated with their possible translations ('word senses'), and clusters of surrounding concepts were used for disambiguation.
The current approach is similar but does *automatic* context disambiguation from *parallel* corpora, and uses only sentential contexts.
4. There are approaches to do disambiguation at the target side, not at the source side. Jassem et al. (2000) use context vectors for translation disambiguation, built on the target side (like in SMT). Target side disambiguation is the current paradigm in SMT (Koehn 2010), and also tried in METIS-II (Carl 2007). This approach must carry all possible transfers of all source words into the target, to disambiguate there. This creates a massive overhead; it could be reduced by using source-language information.

3 Implementation

For terminology, the following section refers to '*transfer*' (or '*entry*') as describing a tuple of <source-lemma, source-POS, target-lemma, target-POS>; it refers to

[2] although the two cases are not clearly distinguished in this work.
[3] „The essence of the proposal is to restrict a word to restrict a word sense inventory to those distinctions that are typically *lexicalized cross-linguistically*"‖ (p. 84).
[4] A similar approach towards transfer can be found in Brown et al. (1991), but they use just *one* contextual 'informant'.

'*package*' as describing a set of transfers sharing the same <source-lemma, source-POS> information. The objective of transfer selection is to find the 'best' transfer out of a package.

3.1 Resource creation

The data for the classifiers were created in the training phase, using on a lexicon, and a bilingual corpus. As transfer selection is language-direction specific, German into English was used in the experiments as direction.

3.1.1 Data

Lexicon

A lexicon was taken as it is used for human lookup: The LinguaDict German>English lexicon [5] comprises 145K German terms and 213K English translations, about 1.5 translations per entry.

First, all 1:1 translations (i.e. packages with only one transfer) were removed from the data set; there is no problem of transfer selection for those entries.

Next, the lexicon was cleaned up by removing function words, entries with differing POS information, and multiword entries.

After these operations, 27.000 packages with 71.400 entries remained for the experiments. Table 3-1 gives the details on the lexicon used for the following analysis.

part of speech	no. packages	no. transfers	no. transfers / package
adjectives	6,900	18,200	2.83
nouns	15,600	35,400	2.27
verbs	4,500	17,800	3.26
total	27,000	71,400	2.63

Tab. 3-1: Packages in the lexicon

A short inspection of the lexicon entries reveals that conceptual transfer alone will never have full coverage: A multitude of transfer selection strategies is required to do proper transfer, as many transfers will not be able to be disambiguated on a purely conceptual level:

- **locale**: (de) '*geschmack*' -> '*flavor*' (en-us) / '*flavour*' (en-uk));
- **spelling**: (en) '*adaptable*' ->: (de-old) '*anpaßbar*' and (de-new) '*anpassbar*';
- **register**: (en) '*adiposity*' -> (de-lit) '*Adipositas*' and (de-coll) '*Verfettung*'.

However, conceptual transfer selection would still be the most frequent selection strategy for this lexicon.

Corpus

The corpus used for the experiments consisted of parallel sentences collected from different domains: Europarl, JRC, news, health&safety, automotive, and others. In total, 3.8M parallel sentences German-English were used.

3.1.2 Corpus Processing

Corpus collection

The corpora were format-converted, lemmatised and tagged (Thurmair et al. 2012) on source and target side; all lemmata were indexed, and for all transfers of each package, a subcorpus of parallel sentences was built which contained the source part (source-lemma, source-POS) and the target part (target-lemma, target-POS) of this entry. For many entries, no such sentence pairs were found; such entries were removed. This operation left 8.12M contexts for relevant term pairs.

Word alignment

In order to avoid accidental co-occurrence of a SL-TL pair, the subcorpora were word aligned, using GIZA++[6]. This operation removed 40% of the contexts, leaving 4.90M sentences, and about 66% of the packages, where no parallel context could be found for any transfer. Table 3-2 shows the remaining data sets.

part of speech	original packages	after bilingual indexing	after word alignment
adjectives	6,900	4,670	1,240
nouns	15,600	11,360	3,690
verbs	4,500	3,930	1,680
Total	27,000	19,960	6,610

Tab. 3-2: Data sets (packages) at the beginning, after bilingual indexing, and after word alignment.

Only 6,600 packages out of the original 27,000 remained for the experiment. So, even in a large parallel corpus, for only 25% of the entries, parallel data can be provided for contextual transfer selection.

About 1000 were subtracted to be used as a test set; the rest was used for classifier building.

3.1.3 Creation of the resources for the classifier

As only for one third of the translation entries, contexts from subcorpus collection would be available, additional information had to be used for the other entries.

dichtung	No	seal	No	integriert|0.33, einreichen|0.16, absaugung|0.16, wirksamkeit|0.16, erforderlich|0.16, einzelheit|0.16, unversehrtheit|0.16, einzeln|0.16, drücken|0.16, einfuellstutzen|0.16, dichtung|0.16, fest|0.16, schließstellung|0.16, verschluss|0.16, gehalt|0.16, position|0.16, sicher|0.16, ausliefern|0.16, einbaufertig|0.16, fetten|0.16
dichtung	No	gasket	No	sem|1.0, semicarbazid|1.0, verschlossen|1.0, dicht|1.0, sterilisation|1.0, (2)|1.0, treibmittel|1.0, kunststoffdichtung|1.0, metalldeckel|1.0, schließe|1.0, glasgefäß|2.0, verwenden|2.0, neu|1.0, erkenntnis|1.0, zerfallen|1.0, azodicarbonamid|2.0, hitzeeinwirkung|1.0, herstellung|2.0, aufgeschäumt|1.0
dichtung	No	packing	No	
dichtung	No	poetry	No	einführung|0.5, werk|0.5, bringen|0.25, gang[|0.25, zivilgesellschaft|0.25, russisch|0.25, spielen|0.25, katalysatorenrolle|0.25, internet|0.25, massenmedien|0.25, buchdruck|0.25, wichtig|0.25, am|0.25, kapitalismus|0.25, finanzinstrument|0.25, belletristik|0.25, osmanisch|0.25, erbe|0.25, klassisch|0.25, islamisch|0.25

Fig. 3-1: Conceptual lexicon: features for the translation of '*dichtung*' ('*seal*', '*poetry*', '*gasket*', '*packing*')

[5] http://www.linguatecapps.com/linguadict

[6] This was possible as the lexicon was restricted to only single word entries.

Therefore a strategy was adopted which is based on two kinds of information:
1. *Conceptual context* clusters, as the original approach suggested. Contexts are just sentences. These data are collected in a conceptual lexicon (*ConcLex*);
2. Translations based on *frequency* information as a fall-back: In case no concept cluster is available, different probability measures are used for transfer selection. They are collected in a probability lexicon (*ProbLex*).

Conceptual Lexicon
1. The **conceptual lexicon** is the resource for the classifier. Each entry gives the following information: 1. The *source* term for which the classifier is called; 2. the *target* term to be selected if this class (translation) is selected; 3. the *features* to be used for the respective class, consisting of pairs of <lemma, probability>, cf. fig. 3-1.
For building the classifier features, simple co-occurrence was computed, related to all words of the respective package. Experiments to optimise the features (restrictions in size, defining thresholds, using distance information, weighting the features, etc.), had no significant effect on the result.

Probability lexicon:
In case a translation had no example sentences, its conceptual cluster was empty; this holds for a significant number of packages, as shown above. Therefore a fall-back strategy was implemented, consisting in a probability computation.
Previous experiments had shown that only using the (target monolingual) *corpus frequency* of a translation is not the best option: We want to know how often the target lemma occurs *as translation of a given source lemma*. Otherwise target lemmata which are very frequent overall (like '*be'* or '*have'*) disturb the transfer selection. So three scores were provided:

- *Package* probability: probability of a given translation related to the other translations of this *package*;
- *Target* probability: probability of a given transfer related to *other source terms* (i.e. for how many SL lemmata is this a possible transfer?)
- *Corpus* probability: probability that this translation is used *at all* in the *target language*. This is the easiest to compute but the least accurate score.

The format of an entry in the probability lexicon is: <source-lemma, source-pos, target-lemma, target-pos, package-prob., target-prob., corpus-prob>.

3.2 Runtime component for transfer selection

The runtime component takes a (source language) translation candidate and a local context (sentence, paragraph). It returns the 'best' transfer for this candidate. Internally, the component first calls the classifier; this was implemented as a naïve Bayes' classifier, using the features of the conceptual lexicon described above. It scores all possible translations of a given source candidate; the best-scoring transfer is returned.
If no classifier is available, the probability lexicon is queried. This is done sequentially, i.e. if a score is zero then the next 'weaker' score is taken. If no probability is available then a random selection is returned.

4 Evaluation

The transfer selection component is tested by determining the translation of a candidate in a given sentential context, and comparing it with the translation used in a reference[7].

4.1 Evaluation criteria

As the LinguaDict lexicon contains many near translations, which can hardly be distinguished on the basis of conceptual transfer, a special evaluation procedure was adopted, consisting of three ranks, instead of a binary 'same/different' decision:
Rank 1: the translation proposed by the system is *identical* to the one in the reference sentence
Rank 2: the proposed translation is close / *similar* to the one in the reference sentence. This was decided to be the case if (a) the proposed translation belongs to the same WordNet synset as the reference; or if (b) the proposed translation is orthographically similar to the reference (like: '*electric*' vs. '*electrical*', '*agglutinating*' vs. ‚*agglutinative*', '*dialogue*' (UK) vs. '*dialog*' (US)).
Rank 3: the two translations are (still) *different*.
The evaluation procedure would accept rank1 and rank2, and reject rank3 results.

4.2 Test data

4.2.1 Test corpus

From all packages where every translation contained more than 5 example sentences, one test sentence was taken, from nouns (700 sentences), verbs (200 sentences), and adjectives (150 sentences); overall the test corpus consisted of about 1000 sentences. The test sentences were not cleaned, just left as in the training corpus.
Each test input is a pair of <source lemma + POS, source sentence context>.

4.2.2 Resources for ranking

For determining rank 2 (similarity), two additional resources were produced:
1. an indexed version of **WordNet** V3[8] , whereby for a given input lemma a list of possible synonyms was retrieved (i.e. the synset lemmata[9]). To do this, all synset lemmata were indexed to the synsets they occur in; this index was later looked up.
It should be noted that WordNet covers the LinguaDict entries only partially (and vice versa); WordNet has 155,200 different entries (including multiwords) while LinguaDict has 210,000 transfers, and 136,000 different English lemmata; but the two resources have only 45,200

[7] Note that evaluation refers to *words*, not to sentences. Therefore procedures like BLEU, TER etc. cannot be used.
[8] http://wordnet.princeton.edu/wordnet/
[9] As the test lexicon contains only single words, also only the single words of the synsets were taken.

entries in common. This fact may influence the evaluation results.

2. a resource for **orthographic similarity**. (a) For all parts of speech, a resource was created which unifies US and UK spelling. (b) For adjectives, additional patterns were considered, like

adj + -ed: ('*abstract*' vs. '*abstracted*')
adj-ic + al: ('*acoustic*' vs. '*acoustical*')
adj+able + ive: ('*adaptable*' vs. '*adaptive*')
adj+ated + ative: ('*agitated*' vs. '*agitative*')
adj + -ous: ('*amorph*' vs. '*amorphous*')

The test frame applies pattern matching for these strings, and simple lookup for the differences in locale.

4.2.3 Test frame

As it was not possible with the available resources to integrate the Xfr component into a complete MT system, a special test system was written which takes a translation candidate (source lemma and POS) and a sentential context as an input, and returns the 'best matching' transfer (target lemma and POS) for this context. This target lemma is then compared to the reference translation, and automatically ranked as explained above: Identical (rank 1) – similar (both words in the same WordNet synset, or orthographically similar) (rank 2) – different (rank 3).

4.3 Test systems

Two test systems were built:

1. one with the full component (called *Xfr-full* below), using both the conceptual and the probability lexicon;
2. one with only the fall-back (called *Xfr-frq* below), using only the probability lexicon.

For each part of speech, a separate run was made, to see if there are significant differences in transfer selection for different parts of speech.

4.4 Test results

The output of the two Xfr systems was first evaluated against the reference translation (absolute evaluation), and then against the output of other MT systems (comparative evaluation).

4.4.1 Absolute evaluation

The test sentences were analysed using the test frame, and the resulting transfer proposal was compared to the reference translation. This was done for both system variants. Baseline is a random selection of transfers. Results are given in Tab. 4-1.

It can be seen that overall 60% of the test terms are translated like in the reference (rank 1), and if similar translations are also taken into account (rank 1+2), then 75% of the test sentences return a correct transfer. All parts of speech show improvements, verbs improving most.

It can also be seen that the conceptual lexicon has a significant effect on the transfer selection; it improves transfer selection by 9% on average, compared to only frequency-based selection (*XFR-frq*), from 66.9% to 75.6%, again with most effect in case of verbs (11%)[10]. In total, two thirds of all transfers were selected based on their conceptual context, the rest is selected based on the frequency fall-back.

	Xfr-full	Xfr-frq	Base line	Absol. Impr.	Relat. Impr.
Noun					
rank1	61.2	49.3	41.6	19.6	47.0
r.1+2	75.2	66.7		33.6	80.6
Adj.					
rank1	58.6	49.7	43.1	15.2	34.9
r1+2	71.7	66.2		28.6	65.0
Verb					
rank1.	61.3	50.5	39.0	22.0	56.2
r.1+2	79.4	68.1		40.4	102.5
Total					
rank1	**60.9**	**49.6**	**41.4**	**19.6**	**47.6**
r.1+2	**75.6**	**66.9**		**34.3**	**83.2**

Tab.4-1: Percentage of correct transfers (rank 1, rank 1+2), and baseline comparison

As a result, the Xfr-full system improves over the baseline by absolute 34%, and relative 83%; improvement is most significant for verbs (with more than 100% relative). For the fall-back system (only frequency-based), the improvement is still 25.7% absolute, and 62.3% relative.

4.4.2 Comparative Evaluation

As no current MT system uses the baseline of a random transfer selection, an additional evaluation step was made to assess the relevancy of the improvement. In order to compare the results with the state of the art, the test sentences were translated with several available MT systems, one SMT and four RBMT systems[11]. The test set was translated with these systems, their translations for the test words were identified and compared to the reference translation.

	Xfr-full	Xfr-frq	SMT	RMT 1	RMT 2	RMT 3	RMT 4
Noun							
r.1	**61.2**	49.3	*55.3*	40.2	44.2	37.9	38.0
r.1+2	**75.2**	66.7	*69.3*	57.8	61.4	55.3	55.8
Adj.							
r.1	**58.6**	49.7	*53.1*	42.8	40.0	40.0	37.2
r1+2	**71.7**	*66.2*	64.1	55.9	54.5	60.0	53.1
Verb							
r.1	**61.3**	*50.5*	47.5	45.1	44.6	33.8	38.7
r1+2	**79.4**	*68.1*	66.2	63.2	65.7	60.3	64.2
Total							
r.1	**60.9**	49.6	*53.5*	41.5	43.7	37.4	38.1
r1+2	**75.6**	66.9	*68.0*	58.6	61.3	57.0	57.0

Tab. 4-2: Comparative evaluation (rank 1 and 1+2) with 1 SMT and 4 RMT systems: % correct selections

[10] This would also be the difference e.g. if the frequency-based fall-back system was used as a baseline.

[11] The systems used were: Google (online version of Dec-2012), Linguatec, Lucy, ProMT, Systran, also with commercial versions available in Dec 2014

Like for the absolute evaluation, *identical* (rank1) and *similar* (rank2) translations were identified. Tab. 4-2 shows the evaluation result.

It can be seen that the *Xfr-full* system clearly shows the best performance of all systems in all categories. It has much better scores than all RMT systems, and also better scores than the SMT. It is absolute 20% better than the least-performing MT system, and still 7% better than the best-performing one.

Even the fall-back frequency-based (*Xfr-frq*) version outperforms all RBMT systems, and is better than the SMT in three of six categories.

Of course there are many parameters which influence this result:

1. The reference translation of the test sentences may not be the best translation, and the test set also contains errors.
2. As for the ranking criteria, not all synonyms (improving from rank 3 to rank 2) are covered by WordNet, esp. as synonyms are context-dependent. So, not all translations which are evaluated as *different* are wrong.
3. As for the RBMT comparison, nothing is known about the transfer lexica used by the other RBMT systems (size, structure etc.), so a real comparison is difficult to make.
4. For SMT, nearly all test sentences (from Europarl etc.) are already in SMT's training set.
5. Many of the lexicon entries tested do not have classifier data, due to data sparsity, even in a 3.8M sentence parallel corpus.

However the result shows that significant improvement in transfer selection can be achieved with the techniques described here, compared to the state-of-the-art MT systems.

5 Assessment

5.1 Relevance

The Xfr approach has the following features:

1. It fits to the architecture of rule-based systems as it provides transfer selection tests on the *source* side, not on the *target* side.
2. The approach is *independent of the specific system structure*, the type of analysis, syntactic structures etc.; it can support shallow MT systems just as well as all kinds of deep RBMT, as it provides a static resource which can easily be linked to any system.
3. For the same reason, it is *more robust* than current selection strategies, which usually fail in cases where the required structure is not built.

5.2 Quality

The quality of the component crucially depends on the quality of the match between the text context and the features of the clusters of the conceptual lexicon. Several options can improve this matching:

1. Extension of the conceptual context for the classifier from sentences to *paragraphs*. This step can improve transfer quality to a level of 96% accuracy (Thurmair 2006). However, most of the parallel data available today are aligned on sentence level, not on paragraph level, so such an approach would be difficult to train.
2. Manual inspection and correction of the clusters, to increase their accuracy.
3. Collection of additional missing concepts by adding *monolingual* correlation analysis to the bilingual one done here.
4. The most serious fact is data sparsity: Only for a fraction of all lexicon entries, corpus data were available, even in large corpora. Therefore an option must be foreseen to have the conceptual lexicon edited by human coders; this would require a review of the current scoring mechanism in the classifier.

5.3 Extensions

To stabilise the results of the current investigation, the following items could be considered:

1. The analysis used only a subset of the lexicon; multiword entries and entries with differing part of speech need to be considered as well..
2. Improve cluster building by collapsing transfers which are clearly synonyms, or variants of each other (like spelling variantsns), *before* the analysis rather than afterwards (in ranking), i.e. using an extended normalisation component in corpus analysis. This can provide more data for such words in clustering.

6 Acknowledgements

The work reported here was carried out by V. Aleksić, Chr. Schwarz, and Gr. Thurmair.

7 References

Aleksić, V., Schwarz, Chr., Thurmair, Gr., Prokopidis, Pr., Papavassiliou, V., Toral, A., 2012: Task-based Evaluation of the PANACEA Production Chain. Panacea Report D8.3

Apidianaki, M.: 2008: Translation-oriented Word Sense Induction based on Parallel Corpora. Proc. LREC Marrakech, Morocco

Brown, P., Della Pietra, St., Della Pietra, V., Mercer, R., 1991: Word-sense disambiguation using statistical methods. Proc. 29th ACL

Carl, M. (2007). METIS-II: the German to English MT system. In Proc. MT Summit XI. Copenhagen, Denmark.

Caseli, H.M., Nunez, M.V, Forcada, M. (2008). Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. In Machine Translation, 20:227–245.

Eberle, K., 2008: Integration von Regel- und Statistik-basierten Methoden in der Maschinellen Übersetzung. In: Seewald-Heeg, U., ed., Maschinelle Übersetzung - von der Theorie zur Anwendung, JLCL 24(3) , Berlin, 2008

Hannemann, Gr., Ambati, V., Clark, J.H., Parlikar, A., Lavie, A. (2009). An Improved Statistical Transfer System for French-English Machine Translation. In Proc. 4th WMT.

Ide, N., Erjavec, T., Tufis, D. (2002). Sense Discrimination with Parallel Corpora. In Proc.

SIG-LEX/SENSEVAL Workshop on Word Sense Dis-ambiguation, Philadelphia, ACL.
Itagaki, M., Aikawa, T., He, X., 2007: Automatic validation of terminology translation consistency with statistical method. Proc. MT Summit Copenhagen
Jassem, K., Graliński, F., Krynicki, Gr. (2000). POLENG – Adjusting a Rule-based Polish-English Machine Translation System by Means of Corpus Analysis. In Proc. 5th EAMT
Koehn, P. (2010). Statistical Machine Translation. Cambridge University Press
Lavie, A., Parlikar, A., Ambati, V. (2008). Syntax-driven Learning of Sub-sentential Translation Equivalents and Translation Rules from Parsed Parallel Corpora. In Proc. 2nd ACL Workshop on Syntax and Structure in Statistical Translation

Menezes, A., Richardson, St. (2001). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In Proc. Workshop on data-driven Machine Translation, ACL 2001, Toulouse
Miháltz, M. (2005). Towards a hybrid approach to word sense disambiguation in Machine Translation. In Proc. RANLP.
Pecina, P., Toral, A., van Genabith, J.,2012: Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation. Proc. COLING Mumbai
Probst, K. (2005). Learning Transfer Rules for Machine Translation with Limited Data. PhD thesis, Carnegie Mellon University.
Probst, K., Carbonell, J., Levin, L. (2002). Semi-automatic learning of transfer rules for machine translation of low-density languages. In Proc. ESSLI.
Resnik, P., Yarowsky, D. (1997). A perspective on Word Sense Disambiguation Methods and their Evaluation. In Proc. ACL SIGLEX workshop on tagging text with lexical semantics: Why, what,and how.
Sánchez-Martínez, F., Forcada, M.L. (2007). Automatic induction of shallow-transfer rules for open-source machine translation. In Proc.11th TMI Conf.
Sánchez-Martínez, F., Forcada, M.L. (2009a). Inferring Shallow-Transfer Machine Translation Rules from Small Parallel Corpora. In Journal of Artificial Intelligence Research 34, 605-635.
Specia, L., Das Graças Volpe Nunez, M., Castello Branco, R.G., Stevenson, M. (2006). Multilingual versus Monolingual WSD. In Proc. Workshop Making Sense of Sense
Thurmair, Gr., 1990: Complex lexical transfer in METAL. Proc. 3rd TMI Conf., Austin, Tx
Thurmair, Gr., 2006: Using Corpus Information to Improve MT Quality. In Proc. Workshop LR4Trans-III, LREC Genova
Thurmair, Gr., Aleksić, V., 2012: Creating term and lexicon entries from phrase tables. Proc. EAMT, Trento
Thurmair, Gr., Aleksić, V., Schwarz, Chr., 2012: Large scale lexical analysis. Proc. LREC, Istanbul
Tyers, F.M., Sánchez-Mártinez, F., Forcada, M.L., 2012: Flexible finite-state lexical selection for rule-based machine translation: Proc EAMT Trento