

A Survey of Domain Adaptation in Machine Translation Towards a refinement of domain space

Lars Bungum, Björn Gambäck
Department of Computer & Information Science
Norwegian University of Science and Technology
Trondheim, Norway

Abstract—Domain adaptation is a recurring problem in Artificial Intelligence in general and Machine Translation (MT) specifically. A system crafted to deal with one particular type of problem often fails when subjected to another, even a closely related one. In MT this is manifested by a system's inability to translate different types of texts (from different domains) with the same confidence. The paper addresses the characterization of language domains and their treatment in the MT literature. We will visit classical linguistic theory as well as cognitive linguistics, and suggest that a refinement of the types of domains along more dimensions can be useful when simultaneously adapting to multiple domains.

1. Introduction

Traditionally, Machine Translation (MT) systems have been designed for one particular language pair, for one specific usage situation, and for one specific application domain. To make such a system more widely applicable, it then has to be generalised along all these dimensions. This paper is primarily concerned with the third one, the issue of domain adaptation in Machine Translation. Two major problem areas are central to this issue.

(i) Firstly, *the properties of a language domain*:

What does it mean that a collection of text belongs to a sub-language, what characterizes it, and accordingly, how can it be automatically identified? Is it merely a question of word-frequency and vocabulary, or do sub-languages also have different syntactical and semantical properties? How is the food domain (e.g., derived from the utterances in a conversation) different from the youth language domain (e.g., derived from the characteristics of the speaker)? We take a special look at Cognitive Linguistics to characterize the domain adaptation problem. Looking at corpus processing seen from the cognitive perspective, we aim to establish a theoretical background upon which a successful system domain adaptation of MT systems can be created. Data-driven approaches have been increasingly influential in the MT literature since the introduction of the IBM models in 1990 [5], and the *usage-based* leanings of cognitive linguistics is a good fit to these wholly usage-based approaches to MT, as the complete training material is taken from real use of language.

(ii) Secondly, *the process of domain adaptation*:

Does it require the transformation of one system into another, or can a system be considered to be *adapted* even if it was designed to work for one specific sub-language from the beginning. And, how can this adaptation process be conducted automatically?

This paper will discuss the treatment of these questions in the literature, showing how *sublanguages* and *domains* are treated in Linguistics. It will go through selected experiments to implement ideas of how to adapt a system designed for one domain to work in another, as well as give an thorough account of the state-of-the-art. We argue that a refinement of language domains and a further investigation of the quantitative sides of (i) will be helpful in answering the challenge poised by (ii). In answering the call for systems to be able to adapt to a particular domain, it is useful to go from adapting to one domain to adapting to many. Rendering one MT system capable of adapting to many domains, henceforth Multiple Adaptation, is an ultimate goal, pending operationalization possibilities identified in the following discussion.

The next section sketches the main characteristics of Machine Translation today. Section 3 looks at the problem of defining what domains are and outline the domain adaptation problem in MT, while Section 4 investigates how domains are treated in Cognitive Linguistics. The core of the paper is Section 5 which discusses how domain adaptation can be achieved and reviews previous efforts to this end. Finally, Section 6 presents the overall findings and ideas for future improvements in domain adaptation.

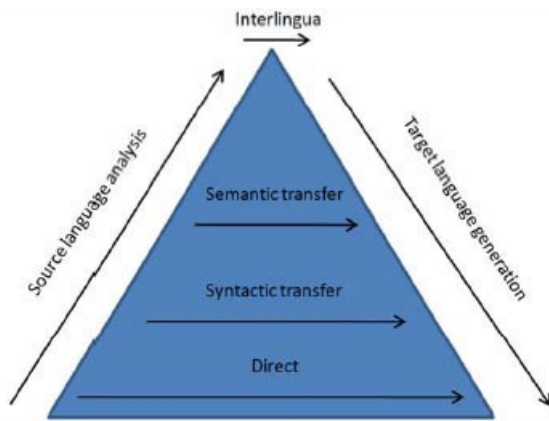


Fig. 1. The Vauquois triangle. Adapted from Vauquois (1976).

2. Machine Translation

The field of Machine Translation (MT), research on using computers to automate the process of translation between human languages (ideas for special-purpose machines are older), ranges back to the early post World War II era when the computers that had been so effective in decrypting codes and calculating bombing patterns were applied to the problem of translation. In early literature, the pioneer Warren Weaver noted that the meaning of individual words depend on the context they are uttered in and hence an important aspect of MT is to determine how much context to include [34]. Weaver's ideas about the prospects for machine translation was received cautiously, in part owing to the complexity of language and the multiple meanings of words. Respect for the complexity of language was voiced right from the beginning of MT history, while initial experiments with success also caused optimism.

Partly due to the infeasibility of quantitative methods due to lack of resources, but also for philosophical reasons, the early attempts at MT were rule-based, at the various levels of interpretation that are characterizable by the Vauquois triangle [33] shown in Figure 1.

Despite the introduction and success of data-driven approaches from the 1990s, rule-based methods are still an important part of the MT family, with an increasing use of hybrid systems, partitioning different jobs of the MT pipeline between data-driven and rule-based methods.

2.1. FAHQMT

In the 1950s, Yehoshua Bar-Hillel coined the phrase FAHQMT, "Fully Automatic High Quality Machine Translation" [1], as the Holy Grail or ultimate goal of MT. Bar-Hillel himself argued the theoretical infeasibility of FAHQMT in an account of the status of automatic translation at that time [2]. By looking at the first two parts

of the acronym as constraints (i.e., 'fully automatic' and 'high quality'), some degree of success could nonetheless be achieved by relaxing them. Involving the user(s) to do pre- and/or post-editing is one way of doing this (hence relaxing the 'fully automatic' constraint); another way is to allow low-quality output from which the user can deduce the main content of the source text and, e.g., decide whether a human translator should be employed to produce a high-quality translation. Controlling the input, or restricting the types of text (to certain language domains) is another way of relaxing the constraints. MT systems have successfully been devised for such limited-domain tasks, especially during the past three decades.

Restricting text to specific domains could be seen as relaxing the 'fully automatic' constraint. Later, however, the acronym has been expanded (in talks by Martin Kay, Xerox PARC) also to comprise 'general purpose', into FAHQGPMT, more explicitly pin-pointing the text type consideration. The purpose dimension says that a system crafted to cater to a language domain has a more specific purpose than a general one. Crafting MT systems fit for special purposes has been important since the dawn of MT, but in recent times the purpose dimension has gotten a new interpretation with the advent of *gisting*, the process of using MT to "get the *gist* of" a text (often a web page), suggesting another refinement of the constraint acronym.

The METEO system [29] is a much-cited success story of MT, restricting the text type to weather data for translation between French and English in bi-lingual Canada. This serves as an example of a system being able to serve one domain. Knowing what it can expect, it is able to fully automatically produce high-quality output.

2.2. Statistical Machine Translation (SMT)

A major paradigm shift in the MT field took place in 1989 when the previously dominant rule-based approaches were challenged by statistical models introduced by IBM [6]. Since then, data-driven MT systems based on statistical inferences from parallel text have been increasing in importance. In statistical approaches, the systems are developed semi-automatically through training on corpora comprised of parallel text, rather than being based on hand-crafted grammar engineering as in the rule-based systems.

SMT systems are comprised of three key parts, a *language model*, based on monolingual data, a *translation model* derived from parallel text, and a *decoder* that finds the best solution from the combination of these two models. The key idea is that the correspondences between two languages are learned from real data, expressed as statistical relations. Examples of such parallel text are the Hansards corpus from the Canadian Parliament and the Europarl corpora from the European Parliament¹.

¹See the Opus project for a collection of free parallel corpora [30].

For language modeling, the degree to which words occur after each other is counted. The frequency of all words (called tokens in corpus linguistics, as not all entities are real words) are counted first, *unigrams*, then how many times they are followed by a specific word, *bigrams*, then how many times they are followed by a specific word, *trigrams*, and so on (counting n-grams). Based on this a model of the language can be expressed in probabilistic terms, giving a probability of a given string belonging to the language. In language modeling, data sparsity is a big problem and receives much attention, as many valid sentences (should not have probability=0) have not been seen before. Unseen words are dealt with primarily through smoothing (taking probability mass from seen events) and back-off (reverting to lower-order n-grams) techniques.

In translation modeling the landscape is more complex, as correspondences between tokens are being extracted. The system then needs to have a notion of what words are a translation of another, a mapping referred to as *alignment*, as the word order and number of words can vary between natural languages. After these mappings have been established, it is possible to count how many times one word is a translation of another creating a model of the translation probabilities. But phenomena like *distortion* (words propensity to move around), *fertility* (how many words in the other language a word will generate) can also be modeled, increasing complexity and training requirements (done with the EM algorithm). By combining alignments in both directions, phrases can also be extracted, that is, larger blocks that consistently translate into larger blocks in the target language. This approach has shown improvement over word-based models.

Finally, when a model of both language and translation is built, this combined model must be *decoded* when faced with a given input string, that is, finding the optimal target language string given the model. Decoding a SMT model has been shown to be NP complete [18], and heuristics are therefore necessary to find a solution within reasonable time. Finding this string is a search problem, and known techniques like A*-search, beam search and other types of pruning are used to make the problem tractable. Here it is possible to introduce heuristics catered to a certain language domain.

3. Language Domains/Sub-Languages

The theoretical discussions in classical Linguistics on text types have mostly been using the term *sublanguages* for specific types of texts within one language. In the MT area, the term *language domain* is more frequent, and both terms will be used interchangeably throughout the text.

Kittredge and Lehrberger [17] outlined the following factors that describe a sublanguage:

- limited subject matter,
- lexical, syntactic and semantic restrictions,
- “deviant” rules of grammar,
- high frequency of certain constructions,
- text structure, and
- use of special symbols.

There work was carried out on the METEO system mentioned in Section 2.1, and details the variations in along the dimensions above. The authors conducted a contrastive study of English and French sublanguages and got results showing strong similarities of the sublanguages within the two languages. Because of these similarities in sentence type and linking devices, the authors remained optimistic towards the automatic translation between sublanguages. They did, however, note that

It should be clear /.../ that a sublanguage is not simply an arbitrary subset of the set of sentences of a language

which elucidates the problem of using the word sublanguage to describe something which is not a subset, because of the apparent connotations.

3.1. Sublanguages and MT

In his 1980 essay “The Proper Place of Men and Machines in Language Translation” [16] Martin Kay maintained that MT only can produce useful results under very special circumstances, echoing the concerns of Bar-Hillel 20 years earlier.

When Kay’s paper was republished in 1997, it was discussed by Alan Melby who pointed towards progress in identifying these special circumstances, by addressing two contrasts in MT input and output [21]. First, there is a contrast between a controlled domain-specific language at one end, and dynamic general language at the other, which is the focal point of this survey. Second, the poles high-quality and indicative translation are two ends of a different spectrum. Melby went on to define a sublanguage as

A sublanguage could be considered to be a case of domain-specific language that is naturally rather than artificially controlled

Plotting these dimensions in a matrix, Melby offered a refinement of the observations of Kay, resulting in a more nuanced view of the state and merits of MT, claiming that it does well in the high-quality, domain-specific box.

3.2. Registers and Sublanguages

In a comment on terminology, Karlgren [14] voiced concerns about using the term sublanguage, and instead proposed the term *register*, borrowed from sociolinguistics. Expanding on the observations by Kittredge and Lehrberger, Karlgren went on to expand on the problematic sides of the mathematical readings of sub-

(and accordingly super-languages) that are not present in natural language. Often sublanguages will exhibit properties that are not present in the perceived *superlanguage*, the general language from which a sublanguage is divided. Registers are defined as a variety of language according to use, in contrast to varieties according to speaker or geographical location. This definition proposed by Karlgren is therefore more process-oriented, to be understood as properties of the speaker directly.

With regard to domain adaptation efforts in MT, such a distinction between sublanguages and registers is useful in order to be specific and clear about what *domain* or *sublanguage* is being discussed.

3.3. Domain Adaptation and SMT

When crafting a SMT system, the training as well as test data can be from different language domains. It has been proven on numerous occasions that a system trained on general text, performs badly on specific domains. Domain adaptation can be done in all three parts. For the monolingual and parallel corpora, the training material can come from different domains, and for the decoding phase, different heuristics can be used to find a solution. As general purpose systems improve due to more training data being made available, the process of domain adaptation — problem (ii) of Section 1 — arises. Building new systems for the specific domains might be very resource-demanding, or even impossible.

4. Cognitive Linguistics

The Cognitive Sciences try to characterize the architecture of thought, the study of our mind and its processes. The interdisciplinary field, also encompasses Linguistics. Being more of an enterprise or an approach than a particular theory of language, Cognitive Linguistics (CL) is not directly comparable to specific theories of grammar. It is rather a study of language which belongs to the Cognitive Science paradigm, maintaining that the theories of language must always be psychologically plausible. While the particular objects of study in CL are much similar to mainstream theories of language, with the two main research avenues being cognitive semantics and cognitive grammar [10], it firmly rejects the notion that there are separate processing entities for grammar and meaning attributed to Generative Grammar frameworks. Thus the cognitive theory of grammar depends on a corresponding theory of semantics (meaning) and can not be understood without.

Evans and Green highlight two key commitments by the research field [10, pp. 32–40]: *generalization commitment* and *cognitive commitment*. The first establishes that the research seeks to find common structural principles for all aspects of language processing such as syntax,

semantics and phonology, dismissing the use of separate theories for each. The second says that findings in CL should reflect what is known about cognition in other fields like philosophy and psychology and not go counter to them.

Croft and Cruise on the other hand cite three central hypothesis in their book on Cognitive Linguistics [9]:

- language is not an autonomous cognitive facility,
- grammar is conceptualization, and
- knowledge of language emerges from language use.

Briefly, the first hypothesis says that the storage of linguistic knowledge is essentially the same as other knowledge (as opposed to being a separate innate facility), whereas the second says that grammar is a concept-forming process which is not impossible to account for with aggregations of truth-conditional statements. The third hypothesis simply states that language is learned through our experience and use of language. This inductive account of language is still able to explain the subtleties of even the most advanced grammatical constructions through descriptions of cognitive processes.

The last hypothesis emphasizes the common ground between CL and data-driven approaches to Language Technology and MT in particular, as mentioned in Section 1. The inferences of translation and language models from mono- and bi-lingual corpora (parallel text) is just that, based on the real usage of language. CL leans on an *empiricist* philosophical tradition, contrasting the *rationalist* foundations of structuralist theories of grammar, often dismissing fringe examples (minor phenomena) for the greater good. As very large amounts of authentic examples can be mined from the net today, assumptions and theories in CL can easily be tested against real data. Analogously, the data-driven approaches that gained ground from the 1990s also form an empiricist answer to a rationalist (rule-based) research field.

4.1. Domains in CL

The notion of *domains* in CL differs from the interchangeable terms domain/sublanguage presented above. A *domain* is understood to be the platform of understanding for a specific *profile* — or the concept represented by the term — and represents what is presupposed when encountering this term. An example is the concept *radius* which presupposes the existence of a circle. This notion of domain is identical to *frame* as used in frame-based semantics, and can also be called *base*. The purpose is to establish a framework to account for how concepts only can be understood within their domain. It is effective in explaining polysemy by showing how the same words can profile almost similar objects in different domains. Croft and Cruise claim that the profile-domain distinction is able to explain why some words are untranslatable [9, pp. 19–21]. The profile of

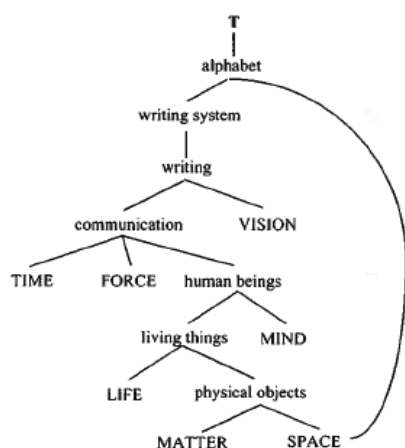


Fig. 2. Domain structure underlying the concept of the letter T. From Croft and Cruse (2004).

the German word *Bildung* can be translated into English, they say, but the domain on which it is understood is so different that the translation is still incorrect. They continue to discuss relations between domains. One profile might correspond to a domain, which in turn is based on other domains. This is exemplified by the domain structure of the letter *T* in Figure 2.

One concept can thus be understood on the basis of many different domains, and not just the immediate domain needed to understand its meaning. The domains are also broadly separated into abstract and base domains. Basic domains are rooted in embodied human experience. Space, time and force are examples of such basic domains. The abstract domains are referred to as the non-basic ones.

4.2. CL Models of Analogy

Gentner and Forbes argue that analogical mapping is a core process in human cognition [11] and present various computational models of analogy, that is, computer programs capable of reasoning about a target domain² by using knowledge of a *base* domain. The process of analogy is divided into four subprocesses [11, p. 267]:

Retrieval: Given a situation, find an analog similar to it.

Mapping: Align the situations structurally, to produce a set of correspondences.

Abstraction: The comparison results may be stored as a schema or other rule-like structures.

Representation: The analogs may be altered to improve partial matches.

Although the authors highlight the insights about human cognition to be derived by computational models of analogy as an end, and furthermore foresee a useful application of analogical reasons in education and

²Here the notion of domain is more like a topic, similar to the one in Section 3.

training, the concrete problem of domain adaptation as described in Section 5 stands out. Still, the structural representation is there also in data-driven MT systems, and the goal of improving the reasoning about the target domain on the basis of another domain is the same.

4.3. Cognitive Linguistics and Corpus Linguistics

Despite the empiricist nature of CL, where all theoretical concepts must be documented in language use, it remained largely non-empirical, until the 1990s when psycholinguistic methods gained ground, and corpus linguistic methods came into play. This paradox surprises Stefanowitsch [28], who takes up a few ideas from Construction Grammar [9, pp. 257–290] and Cognitive Linguistics and investigates how they fare with corpus data. Stefanowitsch shows an example where the behaviour of the verbs *steal* and *rob* are tested for the thematic roles THIEF, TARGET and GOODS, where the goods are robbed/stolen from the target. Using the British National Corpus [8], he shows that *steal* always³ appears with a THIEF and GOODS, whereas *rob* always appears with a THIEF and a TARGET. To a varying degree all three roles will be filled.

The 100M word strong BNC is today considered to be a small corpus, and the idea could be explored in much larger corpora. Such investigations can also be useful in the automatic identification and characterization of domains, which will be elaborated below. Differences in the frequency of ditransitive and dative use of verbs can vary within sub-languages, and other ideas from CL could be used to adapt a base system to a target domain more efficiently.

4.4. Relation to Artificial Intelligence (AI)

The profile-frame/domain distinction from Cognitive Linguistics bears resemblance to frame-based knowledge representation in AI, such as the Protegé system [23]. There separate frames are created as “frames of reference” about concepts. Here it can be explicitly stated that birds can fly, but ostriches not, for example. The database built up can then be queried *based* on the frame.

CL also thinks of the knowledge represented in the frames not as including all real-world instantiations of the frames, but rather as idealized versions of them. With fuzzy borders and categories, it is not the question of whether a word belongs to a certain category, or a phenomenon exhibits some property, but rather their relatedness to it, their *degree of centrality*; how close they are. This is analog to the problems of frame-based knowledge representation with representing conflicting knowledge, and an acknowledgement of that this is sometimes necessary.

³With *always* being defined as “in more than 95% of the cases” — as nothing is considered categorical in actual language use in CL.

5. Domain Adaptation

In the NLP (Natural Language Processing) literature, the second of the two focus areas we stipulated in Section 1 is mostly being addressed, the process of adaptation systems based on text in one domain to text in another. In the WMT 2007 (2nd Workshop on Statistical Machine Translation) shared task, the special challenge was to adapt an SMT system to a specific domain (news) from general text (European Parliament Speeches). In a response to this challenge, Kohn and Schroeder relate to the concrete task, without addressing the formal characteristics of the domains or the adaptation process [20]. Jiang (2008) also defines texts as either being in the *source* domain or in a *target* domain in his review of domain adaptation in statistical classifiers [13]. In experiments with on-line learning and dynamic phrase tables, Sennrich (2011) also devotes little attention to the properties of the Swiss Alpine corpus, and merely describes it as domain specific [26].

Philip Koehn does take up text variation in his book on SMT from 2010 [19]. Koehn explains how words and phrases have different meanings in different domains, and emphasizes two dimensions of text types, namely *modality*, that can be characterized as the purpose of the communication (for instance formal vs. informal text), as well as the text topic, with which the meaning of constituents varies. He explains how the limitation of MT systems to a domain greatly simplifies the task, in accordance with the discussion above, but does not comment on the individual properties along these dimensions of the specific domains he cites.

In a formal analysis, Blitzer [4] addressed domain adaptation in a wider Language Technology context, introducing the notion of structural correspondence learning shared representations between in-domain and out-of-domain data. A *domain* in this formalization is a pair consisting of a distribution \mathcal{D} on \mathcal{X} and a labeling function $h : \mathcal{X} \rightarrow [0, 1]$. Normally two pairs are considered, the *source domain* $\langle \mathcal{D}_S, f_S \rangle$ and a *target domain* $\langle \mathcal{D}_T, f_T \rangle$ — which one attempts to adapt the system to.

5.1. Automatic Identification of Domains

The identification of sublanguages are naturally linked to the identification of languages proper, and to general text classification. Karlgren and Cutting [15] presented an effort to recognize text genres based on discriminant analysis. They separate the process of recognizing a topic from recognizing a genre, though acknowledging the close relation between the two, as some topics may mostly or only be addressed in specific genres.

By aggregating different features from the sub-language texts, they were able to create discriminant functions that could predict if a text belonged to a certain genre based on the parameters learned in the function.

The parameters that were used were counts after a POS-tagger had been employed, of various word classes, but also characters, long words, sentences and the frequency of some trigger words like ‘I’, ‘Me’ and ‘It’. The paper does not explain how genres (or language domains) vary from one another, but outlines a way through which they can be distinguished. Results show that the text genre identification task gets more difficult as the number of categories increase.

Using modern approaches to text classification, extracting metaphors [27] and sentiment [35] across languages, this role can be refined to providing *context* crucial to the disambiguation of intended purpose, as opposed to traditional interaction. Extracting this information about text corpora from various domains can recognize domains in a feature-based model. Later, text classification tasks by means of machine learning have been further developed, using methods such as Maximum Entropy [22], Bayesian Classifiers, or Support Vector Machine learners [32]. These experiments confirm the above-mentioned conjecture that text categories are recognizable using machine learning techniques.

5.2. Automatic Domain Adaptation

Statistical Machine Translation systems are heavily dependent on large data collections and/or annotated data, a scarce and costly resource. The step going from general purpose-intended systems to domain-specific ones can therefore save time for developers and transfer one MT system from poorly-performing in all domains, to well-performing in some, possibly many.

In the context of the language processing duality of rule-based and statistically inspired methodologies, a domain-adapted rule-based system would have to have rules especially crafted to the domain in question, whereas the statistical MT approaches need training data specialized for such a language domain. Annotated data is especially hard to come by and resource-demanding to produce, creating a need for efficient use of such data. Given that a set of samples from the *source* domain and another set from the *target* domain are available, the baseline experiments for creating domain-adapted applications is to combine these sources to create a classifier or a machine translation application capable of handling the *target* domain.

Palmer et al. [24] presented a way of prototyping a MT system based on readily available tools, through the tuning of the system components and their parameters to work better with a specific language domain. The authors pointed to the special syntax in military language, and used a lexical transfer model suited for this. This approach does, however, base itself on the availability of domain-specific training data for use in the process.

Another approach focusing on the components of the MT system is Civera and Juan’s [7] use of *mixture mod-*

eling in the alignment model of the translation system, thereby combining more probability distributions to get better alignment.

Rosenfeld [25] focused on *trigger words* in building adaptive language models. The language models can then be employed in a range of system types such as speech recognizers or MT systems. A Maximum Entropy approach was used to combine the sources of information, so that long-distance triggers (can also be n -grams) combine with the direct n -gram information in the language model. A maximum entropy binary trigger feature was invoked if the trigger had previously been seen in the document given a new word.

Hildebrand et al. [12] continually rebuilt a translation model when encountering a new test sentence, on similar sentences from a training corpus. The similarity scores were based on information retrieval techniques. Essentially, this is a way of refining a more broad training corpus into a domain-specific one after the system has been presented with the sentences to test on.

Tiedemann [31] reflected on a dynamic approach to domain adaptation in statistical MT, after outlining four different paths to follow in domain adaptation, namely

- supervised learning (combination of data),
- unsupervised learning (mixture models),
- better generalizations, and
- dynamic adaptation.

By introducing a way of using a *cache* of new data added from the domain to be adapted to, Tiedemann suggested ways in which parameters can be adjusted on the fly, adapting a running statistical MT system to a new domain using standard tools.

6. Discussion

So far, the theoretical foundations of the *domain adaptation* process have been discussed. We have characterized domains and sublanguages from linguistics and cognitive linguistics, and shown how they are treated in the MT literature. Restricting an application to a specific domain will produce higher quality output. Formally this can be termed as restricting the input to an application, or relaxing some of the three FAHQGPMT ('high quality', 'fully automatic' and 'general purpose') constraints discussed in Section 2.1.

Having a basis for understanding and characterizing how different language domains relate to each other, will make it easier to achieve Multiple Adaptation. Sennrich has reported successful results by weighting different phrase tables built from different sources (corpora of various domains) [26]. After designing a domain matrix over text types and building MT systems for each of them, the different text types can be combined in order to provide the best set-up for an input text at some position in this matrix.

The approaches to domain adaptation described in Section 5.2 have focused on adapting an out-of-domain system to in-domain data, that is, between two types of text. Text domains, however, vary greatly, and a refinement of this domain characterization and labeling may make it easier to adapt a general purpose system to many different domains. By letting the adaptation to specific domains build on each other, a system able to adapt to many domains can be built. Automatic text classification tools can also be used to identify the nature of the input language. We argue that a refinement of the domain matrix will make the job easier, and suggest further research into the subject matter in the following.

From Cognitive Linguistics, we have explored *corpus linguistics* and *analogical reasoning* in Section 4, as well as the building-block nature of domains as a basis of understanding from Croft and Cruise [9]. The quantitative methods described by Stefanowitsch to test CL hypothesis can also be used to describe language domains in the sense of text types. Their (inter)-relatedness and a possible domain phylogeny can be proposed based on the domain cascade idea and verified with quantitative techniques. Finally, analogical reasoning can be used to create adaptations of the same source domain to multiple target domains to achieve Multiple Adaptation.

Mapping different language and translation models, as well as decoding strategies to one another is necessary to make a domain adapting system work, but how to do this remains an open question.

7. Future Work

In further research we will carry out experiments in identifying possible domain phylogenies or domain *matrices*. In the examples from the literature visited in Section 5, the task at hand has been to adapt some system trained on *out-of-domain* data to be able to process *in-domain* data.

We propose devising a phylogeny of domains, using probabilistic weighting of the cascade of domains to be able to save resources in adapting the same system to many different domains. Berg *et al.* employed a similar strategy in their work on grammar induction from related languages [3].

The domains can be categorized along the dimensions visited in the discussion above as modality and topic (which is the prevalent domain labeling), but also alongside other dimensions such as speaker/author age and background, as advances in text classification can identify such text types. Drawing from Cognitive Linguistics, a body of text may belong to several domains at the same time along different dimensions. Analogical reasoning is one way of dealing with multi-dimensional domain spaces, where a body of text can belong to many language domains simultaneously.

Cognitive Linguistics rejects absolute categories and refers to degrees of centrality to an idea. This fits many linguistic phenomena well, and the characterization of sublanguages no less, where a body of text can exhibit properties of many domains at the same time, and in many cases can not meaningfully be characterized as belonging to just one.

Acknowledgments

This work was partially funded by the PRESEMT project (<http://www.presemt.eu>) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number FP7-ICT-4-248307. Thanks to two anonymous reviewers for useful comments on the draft version of the paper.

References

- [1] Y. Bar-Hillel. A quasi-arithmetical notation for syntactic description. *Language*, 29:47–58, 1953. Reprinted in Y. Bar-Hillel. (1964). *Language and Information: Selected Essays on their Theory and Application*, Addison-Wesley 1964, 61–74.
- [2] Y. Bar-Hillel. The present status of automatic translation of languages. *Advances in Computers*, 1:91–163, 1960.
- [3] T. Berg-Kirkpatrick and D. Klein. Phylogenetic grammar induction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1288–1297, Uppsala, Sweden, July 2010. Association for Computational Linguistics.
- [4] J. Blitzer. *Domain Adaptation of Natural Language Processing Systems*. PhD thesis, University of Pennsylvania, 2008.
- [5] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.
- [6] P. F. Brown, V. J. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311, 1993.
- [7] J. Civera and A. Juan. Domain adaptation in statistical machine translation with mixture modelling. In *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation*, pages 177–180, Morristown, NJ, USA, 2007. Association for Computational Linguistics.
- [8] J. H. Clear. *The British National Corpus*, pages 163–187. MIT Press, Cambridge, MA, USA, 1993.
- [9] W. Croft and A. D. Cruse. *Cognitive Linguistics*. Cambridge University Press, 2004.
- [10] V. Evans and M. Green. *Cognitive Linguistics*. Edinburgh University Press, Edinburgh, 2006.
- [11] D. Gentner and K. D. Forbus. Computational models of analogy. *WIREs Cogn Sci*, 2(3):266–276, 2011.
- [12] E. Hildebrand, M. Eck, S. Vogel, and A. Waibel. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, pages 133–142, Budapest, Hungary, May 2005.
- [13] J. Jiang. A Literature Survey on Domain Adaptation of Statistical Classifiers. online, Mar. 2008.
- [14] J. Karlgren. Sublanguages and registers - a note on terminology. *Interacting with Computers*, 5:348–350, 1993.
- [15] J. Karlgren and D. Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International Conference on Computational Linguistics*, volume 2, pages 1071–1075, Kyoto, Japan, Aug. 1994. ACL.
- [16] M. Kay. The proper place of men and machines in language translation. *Machine Translation*, 12:3–23, 1980 & 1997. First appeared as a Xerox PARC Working paper in 1980.
- [17] R. Kittredge and J. Lehrberger, editors. *Sublanguage: studies of language in restricted semantic domains*. W. de Gruyter, Berlin; New York, 1982.
- [18] K. Knight. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607–615, 1999.
- [19] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.
- [20] P. Koehn and J. Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation, StatMT '07*, pages 224–227, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.
- [21] A. Melby. Some notes on the proper place of men and machines in language translation. *Machine Translation*, 12(1/2):29–34, 1997.
- [22] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.
- [23] N. Noy, R. Ferguson, and M. Musen. The knowledge model of protg-2000: Combining interoperability and flexibility. *Lecture Notes in Computer Science*, 1937:69–82, 2000.
- [24] M. Palmer, O. Rambow, and A. Nasr. Rapid prototyping of domain-specific machine translation systems, 1998.
- [25] R. Rosenfeld. A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, 10:187–228, 1996.
- [26] R. Sennrich. Combining multi-engine machine translation and online learning through dynamic phrase tables. In *EAMT-2011: the 15th Annual Conference of the European Association for Machine Translation*, May 2011.
- [27] E. Shutova. Models of metaphor in NLP. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 688–697, Uppsala, Sweden, July 2010. Association for Computational Linguistics.
- [28] A. Stefanowitsch. Cognitive linguistics meets the corpus. To appear in: M. Brdar, M.Z. Fuchs and S.T.Gries (eds) *Converging and Diverging Tendencies in Cognitive Linguistics*.
- [29] B. Thouin. The Meteo system. In V. Lawson, editor, *Practical Experience of Machine Translation*, pages 39–44. North-Holland, Amsterdam, Holland, 1982.
- [30] J. Tiedemann. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, 2009.
- [31] J. Tiedemann. To cache or not to cache? experiments with adaptive models in statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 189–194, Uppsala, Sweden, July 2010. Association for Computational Linguistics.
- [32] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, 2002.
- [33] B. Vauquois. Automatic translation — a survey of different approaches. In H. Karlgren, editor, *Proceedings of the 6th International Conference on Computational Linguistics*, pages 127–135, Ottawa, Canada, 1976. ACL.
- [34] W. Weaver. Translation. In W. N. Locke and A. D. Boothe, editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA, 1949/1955. Reprinted from a memorandum written by Weaver in 1949.
- [35] B. Wei and C. Pal. Cross lingual adaptation: An experiment on sentiment classifications. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 258–262, Uppsala, Sweden, July 2010. Association for Computational Linguistics.