Pictograph Translation Technologies

Vincent Vandeghinste
Dutch Language Institute
Leiden, Netherlands
vincent.vandeghinste@ivdnt.org

Leen Sevens
Centre for Computational Linguistics, KU Leuven
Leuven, Belgium
leen@ccl.kuleuven.be

Ineke Schuurman
Centre for Computational Linguistics, KU Leuven
Leuven, Belgium
ineke@ccl.kuleuven.be

Abstract

We present a set of Pictograph Translation Technologies, which automatically translates natural language text into pictographs, as well as pictograph sequences into natural language text. These translation technologies are combined with sentence simplification and an advanced spelling correction mechanism. The goal of these technologies is to enable people with a low level of literacy in a certain language to have access to information available in that language, and to allow these people to participate in online social life by writing natural language messages through pictographic input. The technologies and demonstration system will be added to the CLARIN infrastructure at the Dutch Language Institute in the course of this year, and have been presented on Tour De Clarin.

1 Introduction

The set of Pictograph Translation Technologies we present consists of Text2Picto, which automatically converts natural language text (Dutch, English, Spanish) into a sequence of Sclera or Beta pictographs, and of Picto2Text, which converts pictograph sequences into regular text (Dutch).

The use of these technologies was instigated by WAI-NOT, a safe internet environment for users with cognitive disabilities, which often also have trouble reading or writing. It was further developed in the EU-funded Able-to-Include project, which built an accessibility layer, allowing software and app developers to build tools that can easily use a number of language technologies, such as the pictograph translation technologies, but also text-to-speech and text simplification. An example of such an application is the e-mail client developed by Saggion et al. (2017).

The Pictograph Translation Technologies for Dutch are further extended in a PhD project in which the tools are not only refined, but also evaluated by a group of targeted users. The initial version of Text2Picto is described in Vandeghinste et al. (2017).

The initial version of Picto2Text is described in Sevens et al. (2015). Refinements consist of the development of a dedicated pictograph selection interface, and of improved translation of pictograph sequences into natural language text through the use of machine translation techniques.

While the current version of the Pictograph Translation Technologies is running on the servers of the Centre for Computational Linguistics at KU Leuven, we are transferring these services to the *Instituut voor de Nederlandse Taal* (Dutch Language Institute), the CLARIN-B centre for Flanders, a region of Belgium which is a member of CLARIN through the flag of the Dutch Language Union (DLU). This transfer will ensure the longevity of the web service, and hence facilitate the ease of communication for people with reading and writing difficulties through the use of this web service beyond the end of the current research projects.

Furthermore, through the extra exposure the service receives as part of CLARIN, we hope to facilitate development of other language technology applications that can use the links between the pictograph sets and the WordNet (Miller, 1995) or Cornetto (Vossen et al., 2008) synsets, as described in Vandeghinste and Schuurman (2014).

A demo of the system and its components can be found at its original location at http://picto.ccl.kuleuven.be/DemoShowcase.html

In what follows we give a brief overview of related work, the system description and the evaluation by the target groups, before we conclude.

2 Related Work

We found only few works related to translating texts for pictograph-supported communication in the literature. Mihalcea and Leong (2009) describe a system for the automatic construction of pictorial representations of the nouns and some verbs for simple sentences and show that the understanding, which can be achieved using visual descriptions, is similar to those of target-language texts obtained by means of machine translation.

Goldberg et al. (2008) show how to improve understanding of a sequence of pictographs by conveniently structuring its representation after identifying the different roles which the phrases in the original sentence play with respect to the verb (structured semantic role labelling is used for this).

Joshi, Wang and Li (2006) describe an unsupervised approach for automatically adding pictures to a story. They extract semantic keywords from a story and search an annotated image database. They do not try to translate the entire story.

Vandeghinste and Schuurman (2014) describe the linking of Sclera pictographs with synonym sets in the Cornetto lexical-semantic database. Similar resources are PicNet (Borman et al., 2005) and ImageNet (Deng et al., 2009), both large-scale repositories of images linked to WordNet (Miller 1995), aiming to populate the majority of the WordNet synsets. These often contain photographs which might be less suitable for communication aids for the cognitively challenged, as they may lack clarity and contrast. The Sclera and Beta pictograph sets are specifically designed to facilitate communication with this user group.

There exist a number of systems that translate pictographs into natural language text (Vaillant 1998; Bhattacharya and Basu, 2009; Ding et al., 2015).[1] Most of these language generation tools expect grammatically or semantically complete pictograph input and they are not able to generate natural language text if not all the required grammatical or semantic roles are provided by the user.

3 System Architecture

Both translation directions make use of the hand-made links between pictographs and Cornetto synsets. Pictographs are linked to one or more Cornetto synsets, indicating the meaning they represent. This has been done for the Sclera and for the Beta set.

3.1 Text2Picto

The first version of this system is described in Vandeghinste et al. (2017). The input text goes through shallow syntactic analysis (sentence detection, tokenization, PoS-tagging, lemmatization, for Dutch: separable verb detection) and each input word is looked up, either in a dictionary (e.g. for pronouns, greetings and other word categories which are not contained in Cornetto) or in Cornetto.

Once the synsets that indicate the meaning of the words in the sentence are identified, the system retrieves the pictographs attached to these synsets. If no pictographs are attached to these synsets, the system uses the relations between synsets (such as hyperonymy, antonymy, and xpos-synonymy) in order to retrieve nearby pictographs. An A* algorithm retrieves the best matching pictograph sequence.

The system was further refined, integrating sentence simplification (Sevens et al., 2017b), as long sequences of pictographs are hard to interpret, temporal detection, as pictograph sequences are usually undefined for morpho-syntacic features and conjugation, spelling correction tuned to the specific user group (Sevens et al., 2016b), which has its own spelling error profile, and proper word sense disambiguation (Sevens et al. 2016a), which identifies the correct sense of polysemous words and retrieves the correct pictograph for that sence.

[1] We do not consider systems that generate the pictographs' labels/lemmas instead of natural language text, or systems that require users (with a motor disability) to choose the correct inflected forms themselves.

3.2 Picto2Text

In the Picto2Text application we have to distinguish the pictograph selection interface from the actual Picto2Text translation engine.

The pictograph selection interface (Sevens et al., 2017a) is a three-level category system. For both Beta and Sclera, there are 12 top categories, which consist of 3 to 12 subcategories each. A total of 1,660 Beta pictographs and 2,181 Sclera pictographs are included, meaning that an average of 21 (for Beta) and 28 (for Sclera) pictographs can be found within each subcategory. The choice for the top-level categories is motivated by the results of a Latent Dirichlet Allocation analysis applied to the WAI-NOT corpus of e-mails sent within the WAI-NOT environment.The following categories were created: *conversation, feelings and behaviour, temporal and spatial dimensions, people, animals, leisure, locations, clothing, nature, food and drinks, objects,* and *traffic and vehicles*. The subcategories were largely formed by exploring Cornetto's hyperonymy relations between concepts. Pictographs occurring within each subcategory are assigned manually. They are ordered in accordance with their frequency of use in the WAI-NOT email corpus, with the exception of logical ordering of numbers (1, 2, 3, ...) and months (January, February, March,...), pairs of antonyms (small and big), or concepts that are closely related. Note that some pictographs can appear in different subcategories.

The Picto2Text translation engine is still under development. The initial system (Sevens et al., 2015) takes a sequence of pictograph names as input, retrieves the synsets to which these pictographs are linked, and generates the full morphological paradigm for each of the lemmas that form that synset. A trigram language model trained on a large corpus of Dutch is used to determine the most likely sequences. In later versions, we are using a fivegram language model trained on a more specific data set, and we are comparing these models with long short-term memory models (LSTM) recurrent neural language models, but have not found improvements yet. A different and promising approach we are pursuing is the use of machine translation tools, such as Moses (Koehn et al., 2007) and OpenNMT (Klein et al., 2017), trained on an artificial parallel corpus, for which the source side (the pictograph side) was automatically generated through the use of the Text2Picto tool, described in section 3.1.

4 Evaluation

Each of the components of the Pictograph Translation Technologies have been evaluated by the users, in two iterations. The first systems have been evaluated through user observations and focus groups. The conclusions of these evaluations were used to make improvements for the second versions, which are currently being re-evaluated. A detailed description of the evaluations is given in Sevens et al. (in press)

5 Conclusions

The Pictograph Translation Technologies, which allow people with reading and/or writing difficulties to participate in the written society are becoming available as a CLARIN tool. These technologies have been developed in such a way that they are easily extendible to other languages and other pictograph sets. They have been developed specifically for users with reading and writing difficulties in mind, but can also be useful for other user groups, in order to resolve communication difficulties, such as migrants that have not learned the language of their host country (yet).

References

[Bhattacharya and Basu 2009] Bhattacharya, S. and Basu, A. (2009). Design of an Iconic Communication Aid for Individuals in India With Speech and Motion Impairments. In *Assistive Technology, n° 21(4)*: 173–187.

[Borman et al. 2005] Borman, A., Mihalcea, R., and Tarau, P. (2005). PicNet: augmenting semantic resources with pictorial representations. In: *Proceedings of the AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors*, pp. 1–7. Menlo Park, California.

[Deng et al. 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, pp. 248–255.

[Ding et al. 2015] Ding C., Halabi N., Alzaben L., Li Y., Draffan E.A. and Wald M. (2015). A Web based Multi-Linguists Symbol-to-Text AAC Application. In *Proceedings of the 12th Web for All Conference.*

[Goldberg et al. 2008] Goldberg, A., Zhu, X., Dyer, C. R., Eldawy, N., and Heng, L. (2008). Easy as ABC? Facilitating pictorial communication via semantically enhanced layout. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL)*, Manchester, England, pp. 119–126.

[Klein et al. 2017] Klein, G., Kim, Y., Deng, Y., Senellart, and Rush, A.M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints* 1701.02810.

[Koehn et al. 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session.* Prague, Czech Republic.

[Joshi et al. 2006] Joshi, D., Wang, J., and Li, J. (2006). The story picturing engine — a system for automatic text illustration. *ACM Transactions on Multimedia Computing, Communications and Applications 2(1):* 1–22.

[Mihalcea and Leong 2009] Mihalcea, R., and Leong, C. W. (2009). Toward communicating simple sentences using pictorial representations. *Machine Translation 22(3)*: 153–173.

[Miller 1995] Miller, G. A. (1995). Wordnet: A lexical database for English. *Communications of the ACM 38(11)*: 39–41.

[Saggion et al. 2017] Saggion, H., Ferrés, D., Sevens, L. and Schuurman, I. (2017). Able to Read my Mail: An Accessible E-mail Client with Assistive Technology. In: *Proceedings of the 14th International Web for All Conference (W4A'17)*. Perth, Australia.

[Sevens et al. 2015] Sevens, L., Vandeghinste, V., Schuurman, I. and Van Eynde, F. (2015). Natural Language Generation from Pictographs. In: *Proceedings of 15th European Workshop on Natural Language Generation (ENLG 2015)*. Brighton, UK. pp. 71-75.

[Sevens et al. 2016a] Sevens, L., Jacobs, G., Vandeghinste, V., Schuurman, I. and Van Eynde, F. (2016b). Improving Text-to-Pictograph Translation Through Word Sense Disambiguation. In: *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*. Berlin, Germany

[Sevens et al. 2016b] Sevens, L., Vanallemeersch, T., Schuurman, I., Vandeghinste, V. and Van Eynde F. (2016a). Automated Spelling Correction for Dutch Internet Users with Intellectual Disabilities. In: *Proceedings of 1st Workshop on Improving Social Inclusion using NLP: Tools and Resources (ISI-NLP, LREC workshop).* Portorož, Slovenia, pp. 11-19

[Sevens et al. 2017a] Sevens, L., Daems, J., De Vliegher, A., Schuurman, I., Vandeghinste, V. and Van Eynde, F. (2017b). Building an Accessible Pictograph Interface for Users with Intellectual Disabilities. In: *Proceedings of the 2017 AAATE Congress*. Sheffield, UK

[Sevens et al. 2017b] Sevens, L., Vandeghinste, V., Schuurman, I. and Van Eynde, F. (2017a). Simplified Text-to-Pictograph Translation for People with Intellectual Disabilities. In: *Proceedings of the 22nd International Conference on Natural Language & Information Systems (NLDB 2017)*. Liège, Belgium.

[Sevens et al. in press] Sevens, L., Vandeghinste, V., Schuurman, I. and Van Eynde F. (in press). Involving People with an Intellectual Disability in the Development of Pictograph Translation Technologies for Social Media Use. In: *Cahiers du CENTAL, Volume 8*. Louvain-La-Neuve, Belgium.

[Vaillant 1998] Vaillant P. (1998). Interpretation of iconic utterances based on contents representation: Semantic analysis in the PVI system. *Natural Language Engineering, n 4(1)*: 17–40.

[Vandeghinste and Schuurman 2014] Vandeghinste, V. and Schuurman, I. (2014). Linking Pictographs to Synsets: Sclera2Cornetto. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland. pp. 3404-3410.

[Vandeghinste et al. 2017] Vandeghinste, V., Schuurman, I., Sevens, L. and Van Eynde, F. (2017). Translating Text into Pictographs. *Natural Language Engineering 23 (2):* 217-244

[Vossen et al. 2008] Vossen, P., Maks, I., Segers, R., and van der Vliet, H. (20080. Integrating lexical units, synsets, and ontology in the Cornetto Database. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pp. 1006–13, Marrakech, Morocco.