Challenge Test Sets for MT Evaluation
– tutorial description –

Maja Popović and Sheila Castilho
ADAPT Centre
School of Computing, DCU
Ireland
{firstname.lastname}@adaptcentre.ie

Abstract

1. **What are challenge test sets (test suites)?**
   - difference between standard test sets and challenge test sets
   - brief overview of the history
2. **Overview of existing challenge test sets**
   - implicit challenge test sets
   - designed challenge test sets
   - testing NMT components
3. **Practical aspects of developing a challenge test set**
   - how to decide what to evaluate
   - how to generate entries
   - how to ensure a straightforward evaluation

1 What are challenge test sets (test suites)?

The main difference between natural (standard) and challenge test sets lies in the distribution of particular phenomena. The traditional ”natural” test sets have uneven distribution of different specific (linguistic) phenomena and therefore are not suitable for getting insight into some particular phenomena. Certain (linguistic or other) phenomena might be under-represented, whereas some others might be over-represented, even repetitive. Challenge test sets (also called ”test suites”), on the other hand, have a good coverage of phenomena of interest, however they do not reflect the statistical distribution of phenomena encountered in naturally occurring data. Therefore, the best strategy for evaluating MT system(s) is to use both natural test sets related to the desired genre/domain as well as challenge test sets covering specific phenomena of interest.

First challenge test sets (CTS) were developed in early 90s for probing syntactic competence of grammar-based MT (and other natural language processing) systems (King and Falkedal, 1990; Way, 1991; Arnold et al., 1993; Lehmann et al., 1996). The emergence of statistical systems suppressed their usage relying mainly on evaluation on standard natural test sets. In the recent years, the idea of using CTS to obtain more fine-grained qualitative observations about MT systems has revived, especially with the emergence of neural machine translation. Nowadays, CTS are not restricted only to syntax, but can cover a broad set of different aspects.

2 Overview of existing challenge test sets

Implicit challenge test sets

A specialised test set can be created implicitly, with no intention to specifically generate a challenging test set. Such test sets are often created as a by-product of evaluation process on standard test sets, for example by evaluating on a subset of a natural test set which contains desired phenomena, such as German compound words (Popović et al., 2006; Escartín, 2012), adjective-noun or verb reorderings (Popović and Ney, 2006), or language features related to gender (Vanmassenhove et al., 2018). These test sets, however, do not necessarily represent well the distribution of the phenomena of interest.

Designed challenge test sets

A number of test sets have been explicitly designed for particular phenomena in the last years, especially since the emergence of the “Additional Test Suites” sub-task in the framework of the WMT translation shared task in 2018.

Some of the first ”revived” test suites cover a large taxonomy with different categories, such as (Burchardt et al., 2017; Macketanz et al., 2018; Avramidis et al., 2019) for German-to-English and (Isabelle et al., 2017) for English-to-French.

Other test sets concentrate on one broad phenomenon and several sub-categories within it, such as grammatical (Sennrich, 2017; Cinková and Bojar, 2018) or morphological divergences (Burlot and Yvon, 2017; Burlot et al., 2018), or discourse phenomena (Šoštarić et al., 2018; Bawden et al., 2018; Bojar et al., 2019; Voita et al., 2019). Another type concentrates on one particular phenomenon and one or more of its potential variations, for example ambiguity of pronouns (Guillou and Hardmeier, 2016; Guillou et al., 2018; Müller et al., 2018; Bawden et al., 2018), lexical ambiguity of nouns (Rios Gonzales et al., 2017; Rios Gonzales et al., 2018; Raganato et al., 2019), ambiguous conjunctions (Popović and Castilho, 2019; Popović, 2019). , or gender bias for nouns related to professions (Stanovsky et al., 2019).

Various language pairs and translation directions have been covered, almost always including English as one of the languages: French (Isabelle et al., 2017; Guillou and Hardmeier, 2016; Popović and Castilho, 2019; Popović, 2019), German (Burchardt et al., 2017; Müller et al., 2018; Rios Gonzales et al., 2018; Burlot et al., 2018; Popović and Castilho, 2019; Popović, 2019), Czech (Burlot and Yvon, 2017; Burlot et al., 2018; Cinková and Bojar, 2018; Bojar et al., 2019; Raganato et al., 2019), Latvian (Burlot and Yvon, 2017), Lithuanian (Raganato et al., 2019), Finnish (Burlot et al., 2018; Raganato et al., 2019), Russian (Raganato et al., 2019; Voita et al., 2019), Serbian, Croatian, Spanish, Portuguese (Popović and Castilho, 2019), Turkish (Burlot et al., 2018).

Some of the test sets are constructed and evaluated fully manually (Isabelle et al., 2017; Burchardt et al., 2017), whereas some others fully rely on automatic generation and evaluation process without any human inspection and intervention (Burlot and Yvon, 2017; Burlot et al., 2018; Raganato et al., 2019; Stanovsky et al., 2019). The majority of the test suites relies, to a greater or lesser proportion, both on automatic and on manual processes.

The majority of the designed challenge test sets are publicly available.

Testing NMT components

The most recent trend of evaluating specific aspects in an NMT model are so called ”probing tasks”. The goal of these tasks is to test the outputs of different neural network components in order to determine whether the (linguistic) information of interest is well captured in the obtained (word, sentence or other) representations. A probing task is a simple classification problem focused on desired phenomena, and requires a task which represents well these phenomena. This classification task needs an appropriate data set which represents well the property of interest and is large enough to be split into training and validation part for the classifier.

Several layers of RNN-based NMT systems have been tested on morphology, syntax and semantics (Belinkov et al., 2017a; Belinkov et al., 2017b; Durrani et al., 2019) using a corresponding tagging tasks (POS, syntactic and semantic tagging). The main findings are that lower layers are better capable of capturing morphology whereas semantics is better represented in higher layers.

3 Practical aspects of developing a challenge test set

Although the CTSs can be quite different depending on the main goals of evaluation, on the amount and types of phenomena they contain, as well as on the languages they cover, they all have certain common practical aspects to be taken into account.

Deciding on phenomena of interest

The decision about what to evaluate exactly is far from trivial. One of the factors is the main goal of the evaluation: understanding a particular phenomenon, understanding a particular system, comparing different technologies, investigating different languages, etc. Furthermore, this decision can be reached top-down (start from theoretically problematic concepts) (Burchardt et al., 2017; Isabelle et al., 2017) or bottom-up (start from concepts which are found to be problematic in natural test sets) (Šoštarić et al., 2018; Voita et al., 2019).

.

Construction process

Once the categories to evaluate are determined, creating entries in an CTS can be performed in different ways: collecting natural data samples which contain desired phenomena, editing and paraphrasing natural entries, or completely inventing ("artificial"). Furthermore, depending on the phenomena and application, some CTS should contain only entries with phenomena of interest, whereas some should also contain contrastive (negative) entries where the phenomena of interest are absent, or entries which contain some errors. Additional point of view is whether the CTS should be model agnostic (suitable for comparison of different technologies), or designed for evaluating development of a specific architecture. Last but not least, the evaluation procedure should be as straightforward as possible, especially if it involves manual inspection.

4 Presenters

Maja Popović

ADAPT Centre, Dublin City University, Ireland
maja.popovic@adaptcentre.ie
https://www.adaptcentre.ie/about/team

Maja Popović is a post-doctoral researcher at the ADAPT Centre at Dublin City University. She graduated at the Faculty of Electrical Engineering, University of Belgrade, Serbia, and continued her studies at the RWTH Aachen University, Germany, where she obtained her PhD with the thesis "Machine Translation: Statistical Approach with Additional Linguistic Knowledge". After that, she continued her research at the German Institute for Artificial Intelligence (DFKI), at the Humboldt University of Berlin, and currently at the ADAPT Centre at Dublin City University. Her research interests include machine translation, automatic and human evaluation in NLP, as well as combining linguistic knowledge and data-driven methods. She published over 60 papers in various conferences, workshops and journals, including a book chapter about machine translation evaluation. She serves as regular programme committee member of ACL/EACL/NAACL, EMNLP, COLING, LREC and other conferences and workshops for more than 10 years. She has been a co-organiser of the Workshop on The Qualities of Literary Machine Translation at MT Summit 2019 and the Workshop on Quality Assessment for Text Simplification at LREC 2016, tutorial chair at ACL 2017 and area chair at EAMT and COLING 2018. She is a guest co-editor of the special issue of Machine Translation Journal on Human Factors in Neural Machine Translation published in 2019 by Springer. She gave several invited talks and lectures about evaluation for machine translation and other NLP tasks.

Sheila Castilho

ADAPT Centre, Dublin City University, Ireland
sheila.castilho@adaptcentre.ie
https://www.adaptcentre.ie/about/team

Sheila Castilho graduated in Linguistics and Education from the UNIOESTE University in Brazil. She holds a joint Master in Natural Language Processing from the University of Wolverhampton, UK and the University of Algarve, Portugal. She completed her PhD dissertation at Dublin City University in 2016. Currently, she is a postdoctoral researcher at the ADAPT Centre. She is a programme committee member of a number of translation, machine translation and NLP conferences and has acted as a reviewer for high-profile journals. She has authored several journal articles and book chapters on translation technology, post-editing of machine translation, user evaluation of machine translation, and translators' perception of machine translation. She is a co-editor of the book 'Translation Quality Assessment: From Principles to Practice', published in 2018 by Springer and a guest co-editor of the special issue of Machine Translation Journal on Human Factors in Neural Machine Translation published in 2019 by Springer. Her research interests include machine translation, post-editing, machine and human translation evaluation, usability, and translation technologies.

5 Acknowledgements

The ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Dublin City University is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Arnold, D., Moffat, D., Sadler, L., and Way, A. (1993). Automatic test suite generation. *Machine Translation*, 8(1):29–38.

Avramidis, E., Macketanz, V., Strohriegel, U., and Uszkoreit, H. (2019). Linguistic Evaluation of German-English Machine Translation Using a Test Suite. In *Proceedings of the 4th Conference on Machine Translation (WMT 19)*, Florence, Italy.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 1304–1313, New Orleans, Louisiana.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2017a). What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 17)*, pages 861–872, Vancouver, Canada.

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., and Glass, J. (2017b). Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP 17)*, pages 1–10, Taipei, Taiwan.

Bojar, O., Rysová, K., Rysová, M., and Musil, T. (2019). Manual Evaluation of Discourse Relations Translation Accurracy in Document Level NMT. In *Proceedings of the 4th Conference on Machine Translation (WMT 19)*, Florence, Italy.

Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.-T., and Williams, P. (2017). A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, 108(1):159–170.

Burlot, F., Scherrer, Y., Ravishankar, V., Bojar, O., Grönroos, S.-A., Koponen, M., Nieminen, T., and Yvon, F. (2018). The WMT'18 Morpheval test suites for English–Czech, English–German, English–Finnish and Turkish–English. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 550–564, Belgium, Brussels.

Burlot, F. and Yvon, F. (2017). Evaluating the morphological competence of Machine Translation Systems. In *Proceedings of the 2nd Conference on Machine Translation (WMT 17)*, pages 43–55, Copenhagen, Denmark.

Cinková, S. and Bojar, O. (2018). Testsuite on Czech–English Grammatical Contrasts. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 565–575, Belgium, Brussels.

Durrani, N., Dalvi, F., Sajjad, H., Belinkov, Y., and Nakov, P. (2019). One size does not fit all: Comparing NMT representations of different granularities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019)*, pages 1504–1516, Minneapolis, Minnesota.

Escartín, C. P. (2012). Design and compilation of a specialized Spanish–German parallel corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 12)*, Istanbul, Turkey.

Guillou, L. and Hardmeier, C. (2016). PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 16)*, Portorož, Slovenia.

Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E., and Loáiciga, S. (2018). A Pronoun Test Suite Evaluation of the English–German MT Systems at WMT 2018. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 576–583, Belgium, Brussels.

Isabelle, P., Cherry, C., and Foster, G. (2017). A Challenge Set Approach to Evaluating Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 17)*, pages 2486–2496, Copenhagen, Denmark.

King, M. and Falkedal, K. (1990). Using Test Suites in Evaluation of Machine Translation Systems. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING 90)*, pages 211–216, Helsinki, Finland.

Lehmann, S., Oepen, S., Regnier-Prost, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvry, F., Estival, D., Dauphin, E., Compagnion, H., Baur, J., Balkan, L., and Arnold, D. (1996). TSNLP - Test Suites for Natural Language Processing. In *Proceedings of the 16th conference on Computational linguistics (COLING 96)*, pages 711–716, Copenhagen, Denmark.

Macketanz, V., Avramidis, E., Burchardt, A., and Uszkoreit, H. (2018). Fine-grained evaluation of German–English Machine Translation based on a Test Suite. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 584–593, Brussels, Belgium.

Müller, M., Rios Gonzales, A., Voita, E., and Sennrich, R. (2018). A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 61–72, Belgium, Brussels.

Popović, M. (2018a). Error Classification and Analysis for Machine Translation Quality Assessment. In Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S., editors, *Translation Quality Assessment: From Principles to Practice*, pages 129–158. Springer, Cham.

Popović, M. (2018b). Language-related issues for NMT and PBMT for English–German and English–Serbian. *Machine Translation*, 32(3):237–253.

Popović, M. (2019). Evaluating conjunction disambiguation on English-to-German and French-to-German WMT 2019 systems. In *Proceedings of the 4th Conference on Machine Translation (WMT 19)*, Florence, Italy.

Popović, M., Arčan, M., and Klubička, F. (2016). Language related issues for machine translation between closely related south Slavic languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 43–52, Osaka, Japan.

Popović, M. and Arčan, M. (2015). Identifying main obstacles for statistical machine translation of morphologically rich South Slavic languages. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 15)*, pages 97–104, Antalya, Turkey.

Popović, M. and Castilho, S. (2019). Are ambiguous conjunctions problematic for machine translation? In *Proceedings of the 12th Conference on Recent Advances in Natural Language Processing (RANLP 19)*, Varna, Bulgaria.

Popović, M. and Ney, H. (2006). POS-based word reorderings for statistical machine translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06)*, Genoa, Italy.

Popović, M., Stein, D., and Ney, H. (2006). Statistical Machine Translation of German Compound Words. In *Proceedings of the 5th International Conference on Advances in Natural Language Processing (FinTal 06)*, pages 616–624, Turku, Finland.

Raganato, A., Scherrer, Y., and Tiedemann, J. (2019). The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation. In *Proceedings of the 4th Conference on Machine Translation (WMT 19)*, Florence, Italy.

Rios Gonzales, A., Mascarell, L., and Sennrich, R. (2017). Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the 2nd Conference on Machine Translation (WMT 17)*, pages 11–19, Copenhagen, Denmark.

Rios Gonzales, A., Müller, M., and Sennrich, R. (2018). The Word Sense Disambiguation Test Suite at WMT18. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 594–602, Belgium, Brussels.

Sennrich, R. (2017). How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain.

Stanovsky, G., Smith, N., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 19)*, Florence, Italy.

Vanmassenhove, E., Hardmeier, C., and Way, A. (2018). Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 18)*, pages 3003–3008, Brussels, Belgium.

Voita, E., Sennrich, R., and Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 19)*, Florence, Italy.

Vojtěchová, T., Novák, M., Klouček, M., and Bojar, O. (2019). SAO WMT19 Test Suite: Machine Translation of Audit Reports. In *Proceedings of the 4th Conference on Machine Translation (WMT 19)*, Florence, Italy.

Šoštarić, M., Hardmeier, C., and Stymne, S. (2018). Discourse-Related Language Contrasts in English-Croatian Human and Machine Translation. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 36–48, Belgium, Brussels.

Way, A. (1991). Developer-Oriented Evaluation of MT Systems. In *Proceedings of the Evaluators' Forum*, pages 237–244, Les Rasses, Switzerland.

Challenge Test Sets for MT Evaluation

Engaging Content
Engaging People

Maja Popović, Sheila Castilho
ADAPT Centre @ DCU
name.surname@adaptcentre.ie

DUBLIN
MACHINE
TRANSLATION
SUMMIT 2019

MT Summit 2019
Dublin City University
19 August 2019

Challenge Test Sets for MT Evaluation

Engaging Content
Engaging People

- material:
  `https://sites.google.com/view/challenge-test-sets-tutorial/home`
  - these slides
  - tutorial description
  - full list of useful references
- questions:
  - after a chapter/topic
  - during the break
  - after the tutorial
  - Twitter #MTSummitCTS

Challenge Test Sets for MT Evaluation

Engaging Content
Engaging People

We'll be live tweeting it!

- #MTSummitCTS - our tutorial hashtag
- @amelija16mp and @_SheilaCastilho - the presenters
- @MTSummitXVII - the conference Twitter account
- #MTSummit2019 - the conference hashtag

Goals

- provide a broad overview of specified "challenge" test sets ("test suites") developed for different aspects of MT evaluation
- provide practical advice for creating own test suites

Note:

- it is not possible to cover all details from all papers
- a list of references is available in the tutorial description

Outline

Engaging Content
Engaging People

1. Introduction
2. Overview of various challenge test sets
3. Practical aspects of designing challenge test sets

Engaging Content
Engaging People

1. Introduction

- What are test sets?
- Standard Test Sets
- Challenge Test Sets - Test Suites
- Test Sets in MT tasks
- A bit of history

What are test sets?

Engaging Content
Engaging People

two types of test sets

- standard test sets
  - sub-sets of naturally occurring data
  - usually of the same nature as the training set
  - if not: out-of-domain
    challenging in a way, but not a challenge test set

- challenge test sets (CTs) - also known as Test Suites
  - specifically designed for particular phenomena
  - not related to the training set
  - not related to the domain

What are challenge test sets - test suites?

Engaging Content
Engaging People

Generally:

- it has to concentrate on specific phenomena
- it should represent well these phenomena
- it should be of reasonable size
- it should enable a straightforward evaluation
- the phenomena are usually linguistically motivated

Therefore, our definition of Challenge Test Sets is:

A challenge test set is a representative set of isolated or in-context sentences, each hand or (semi)automatically designed to evaluate a system's capacity to translate a specific linguistic phenomenon.

What are challenge test sets - test suites?

Fun Fact: "Suites" is pronounced /swiːt/ just like "sweets"!

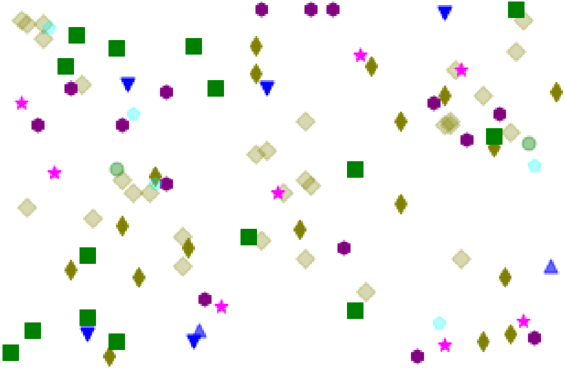
What are challenge test set - test suites?

Engaging Content
Engaging People

Not like "suits"!

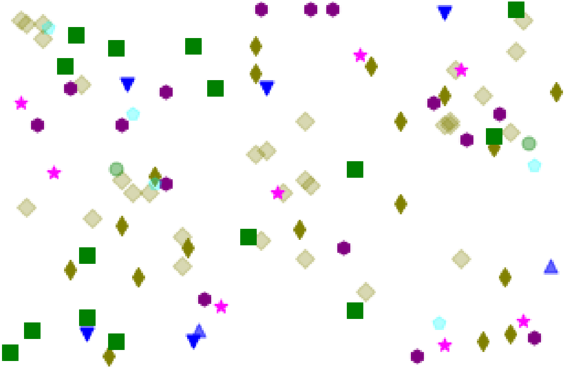
Standard test sets

geometric shapes



Standard test sets

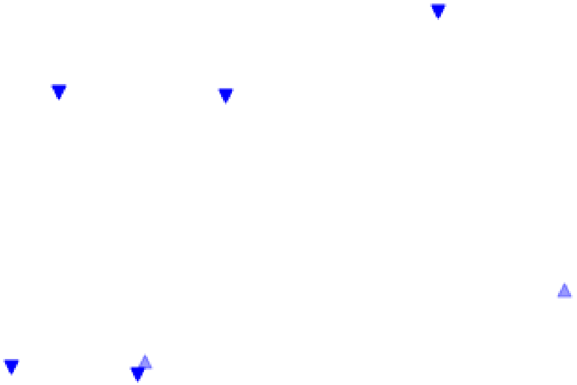
geometric shapes



Performance on triangles?

Evaluate on triangles

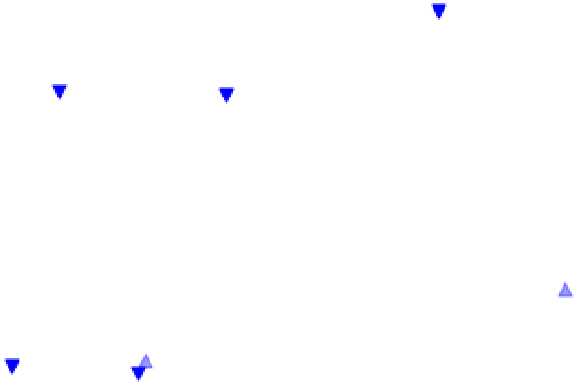
take triangles from the standard test set



Triangles from the standard test set

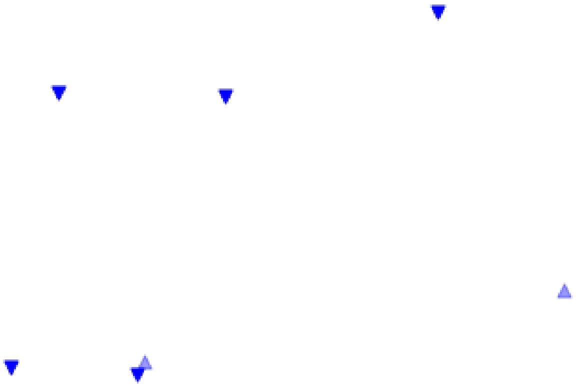
- rarely occurring

Triangles from the standard test set



- rarely occurring
- uneven distribution of two types
  (two upwards and five downwards)

Triangles from the standard test set



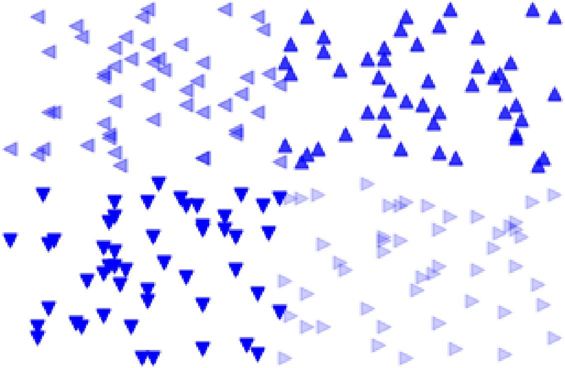
- rarely occurring
- uneven distribution of two types (two upwards and five downwards)
- what about leftwards and rightwards?

Standard test sets

Engaging Content
Engaging People

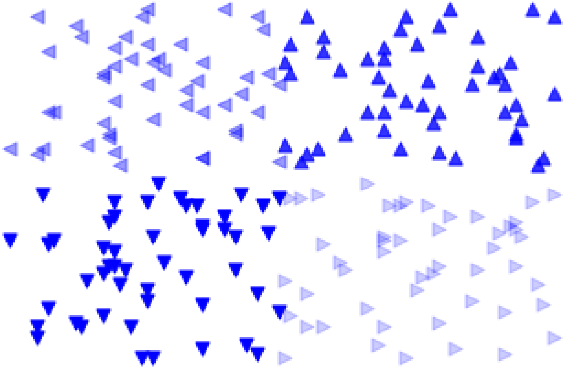
Pretty much like

Challenge test sets: A test Suite for triangles



Challenge test sets: A test Suite for triangles

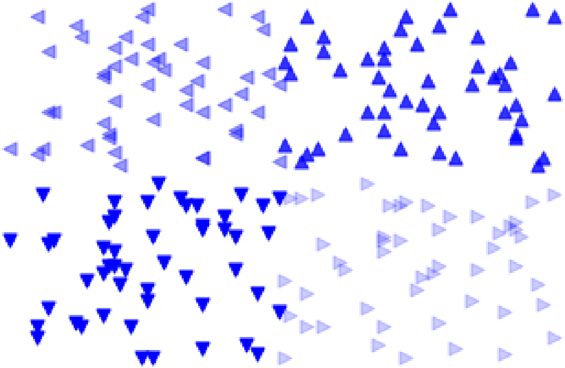
Engaging Content
Engaging People



- balanced distribution

Challenge test sets: A test Suite for triangles

Engaging Content
Engaging People



- balanced distribution
- covers all four types of triangles

Challenge test set - test suite

Pretty much like

Engaging Content
Engaging People

Test Sets in MT tasks

Engaging Content
Engaging People

A part of WMT 2019 standard test set

Shark injures 13-year-old on lobster dive in California
A shark attacked and injured a 13-year-old boy Saturday while he was diving for lobster in California on the opening day of lobster season, officials said.
The attack occurred just before 7 a.m. near Beacon's Beach in Encinitas.
Chad Hammel told KSWB-TV in San Diego he had been diving with friends for about half an hour Saturday morning when he heard the boy screaming for help and then paddled over with a group to help pull him out of the water.
Hammel said at first he thought it was just excitement of catching a lobster, but then he "realized that he was yelling, 'I got bit!
His whole clavicle was ripped open," Hammel said he noticed once he got to the boy.
"I yelled at everyone to get out of the water: 'There's a shark in the water!'" Hammel added.
The boy was airlifted to Rady Children's Hospital in San Diego where he is listed in critical condition.
The species of shark responsible for the attack was unknown.
Lifeguard Capt. Larry Giles said at a media briefing that a shark had been spotted in the area a few weeks earlier, but it was determined not to be a dangerous species of shark.
Giles added the victim sustained traumatic injuries to his upper torso area.
Officials shut down beach access from Ponto Beach in Casablad to Swami's in Ecinitas for 48 hours for investigation and safety purposes.
Giles noted that there are more than 135 shark species in the area, but most are not considered dangerous.

WMT 2019 standard test set: "but" and "and"

Shark injures 13-year-old on lobster dive in California

A shark attacked **and** injured a 13-year-old boy Saturday while he was diving for lobster in California on the opening day of lobster season, officials said.

The attack occurred just before 7 a.m. near Beacon's Beach in Encinitas.

Chad Hammel told KSWB-TV in San Diego he had been diving with friends for about half an hour Saturday morning when he heard the boy screaming for help **and** then paddled over with a group to help pull him out of the water.

Hammel said at first he thought it was just excitement of catching a lobster, **but** then he "realized that he was yelling, 'I got bit!

His whole clavicle was ripped open," Hammel said he noticed once he got to the boy.

"I yelled at everyone to get out of the water: 'There's a shark in the water!'" Hammel added.

The boy was airlifted to Rady Children's Hospital in San Diego where he is listed in critical condition.

The species of shark responsible for the attack was unknown.

Lifeguard Capt. Larry Giles said at a media briefing that a shark had been spotted in the area a few weeks earlier, **but** it was determined not to be a dangerous species of shark.

Giles added the victim sustained traumatic injuries to his upper torso area.

Officials shut down beach access from Ponto Beach in Casablad to Swami's in Ecinitas for 48 hours for investigation **and** safety purposes.

Giles noted that there are more than 135 shark species in the area, **but** most are not considered dangerous.

A challenge test set for "but" and "and"

An additional test – **and** it would not be difficult to prepare – would make the results stronger.
And now something completely different!
And what is she doing here?
Who is she **and** what is she doing here?
Andy got a package **and** Jack got a letter.
Ann likes to dance, **and** Bill likes to dance, too.
Because I had it **and** now I do not have it.
I'm a great actor, **and** you're a cheap producer.
Cathy thought she was going to win, **and** you pushed her.
Chris planned this trick **and** you carried it through.
Come to us not as a guest, **but** as a brother.
Convicted not of arson, **but** of some minor transgression.
Crime is not the reason **but** the consequence.
Do not hate the sinner, **but** the sin.
Do not immediately refuse the man, **but** show what his weak points are.
Don't apologize to me, **but** to her.
Don't talk, **but** do it now.
Dreaming won't get you to your goal, **but** discipline will.
Enlightenment is not found in falling **but** in rising.
Family rituals aren't necessarily about what you do, **but** more about doing things together.

Engaging Content
Engaging People

The main differences

Engaging Content
Engaging People

standard test sets

- + reflect the frequency distribution of different phenomena found in naturally occurring data
- – particular phenomena might be under-represented, and others might be over-represented

challenge test sets

- + have a good coverage of phenomena of interest
- – do not reflect the statistical distribution of phenomena encountered in real corpora

History

- early 1990s:
  evaluating grammar competence of rule-based NLP [Lehmann et al., 1996] or MT [King and Falkedal, 1990, Way, 1991] systems

- early 2000s:
  suppressed by emergence of statistical systems

  probable reason:
  - the performance of statistical systems depends very much on the particular training data and parameter settings
  - conclusions about the grammatical errors they make are difficult to draw

History

- since 2015:
  revived in order to obtain more fine-grained qualitative observations about MT systems

- since 2017:
  expanded with the emergence of neural systems

- since 2018:
  "Additional Test Suites in News Translation Task" at WMT[1],[2]

[1] http://www.statmt.org/wmt18/translation-task.html

[2] http://www.statmt.org/wmt19/translation-task.html

History

- WMT "Additional Test Suites in News Translation Task"

2. Overview of Various Challenge Test Sets (CTSs)

Overview of various CTSs

- first CTSs
- implicit CTSs
- designed CTSs
- probing NMT representations

First CTSs – NLP in general

Engaging Content
Engaging People

TSNLP (Test Suites for Natural Language Processing) Project [Lehmann et al., 1996][3]

- English, French and German
- restricted vocabulary
- as short as possible entries
- each entry focuses only on a single phenomenon
- about 10 types of core phenomena + their sub-classes
- correct (well-formed) and contrastive (ill-formed) entries
  - contrastive = introducing errors
    deletions, substitutions, omissions and reorderings
- evaluation: the system output is either acceptable or unacceptable

[3] http://www.delph-in.net/tsnlp/

First CTSs – MT

Engaging Content
Engaging People

MT test suites for probing RBMT systems [King and Falkedal, 1990]

- Describes a theory for translation strategy using test suites
- test suite based on actual texts
- outputs classified by the evaluator as acceptable or unacceptable
- use two parallel test suites, one to be used by the evaluator to give feedback to the manufacturer (translation problems), and one to serve as a control corpus (language coverage).

First CTSs – MT

Engaging Content
Engaging People

MT test suites for probing RBMT systems [Way, 1991]

- suggestions about the construction of test suites
- test suite based on actual texts
- test suite constructed with information concerning the relative frequency of phenomena obtained from a corpus
- the combinations of concepts need to be limited, to prevent sentences becoming intolerably complicated
- distinction of 4 test suite types:
  - an initial development suite
  - an analysis suite
  - a synthesis suite
  - a transfer suite

Implicit CTSs

interested in system's performance on X?

- take all Xs from your standard test set
- calculate evaluation scores on this sub-set

X={linguistic phenomenon, geometric shape, candy,...}

Implicit CTSs

a sub-set of a standard test set

- created as a by-product of usual evaluation process
- also used for evaluating statistical systems
- do not necessarily represent well the distribution of the phenomena of interest

Implicit CTSs

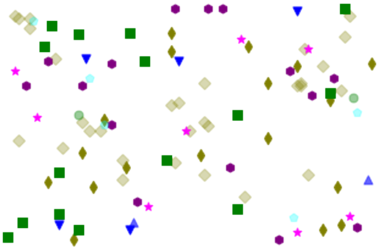
Engaging Content
Engaging People

phenomena

- German compounds [Popović et al., 2006, Escartín, 2012]
- noun-adjective and verb reordering [Popović and Ney, 2006]
- gender [Vanmassenhove et al., 2018]
- certainly some more

Implicit CTSs

standard test set



Performance on triangles and quadrilaterals?

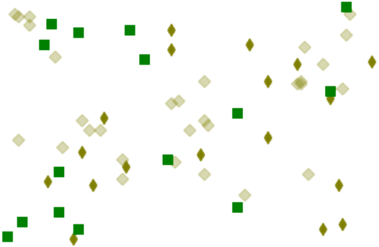
Implicit CTSs

implicit test set for triangles

- not very good

Implicit CTSs

implicit test set for quadrilaterals



- much better

! one never knows what is "hidden" in the standard test set

Implicit CTSs

Engaging Content
Engaging People

"and" and "but" from the standard WMT 2019 test set

A shark attacked **and** injured a 13-year-old boy Saturday while he was diving for lobster in California on the opening day of lobster season, officials said.

Chad Hammel told KSWB-TV in San Diego he had been diving with friends for about half an hour Saturday morning when he heard the boy screaming for help **and** then paddled over with a group to help pull him out of the water.

Hammel said at first he thought it was just excitement of catching a lobster, **but** then he "realized that he was yelling, 'I got bit!

Lifeguard Capt. Larry Giles said at a media briefing that a shark had been spotted in the area a few weeks earlier, **but** it was determined not to be a dangerous species of shark.

Officials shut down beach access from Ponto Beach in Casablad to Swami's in Ecinitas for 48 hours for investigation **and** safety purposes.

Giles noted that there are more than 135 shark species in the area, **but** most are not considered dangerous.

Designed CTSs

Engaging Content
Engaging People

interested in system's performance on X?

- design a controlled test set specifically for X
- specify the evaluation procedure for X
- evaluate on this test set

X={linguistic phenomenon, geometric shape, candy,...}

Designed CTSs

Engaging Content
Engaging People

a number of CTSs has been developed in the last years (from 2016 till mid-2019)

- different amounts and types of phenomena (from large taxonomies to a single phenomenon)
- various language pairs and translation directions (almost always involving English)
- different strategies for generation/evaluation (manual, semi-automatic, automatic)

Designed CTSs

taxonomies

- large taxonomies covering a number of distinct phenomena
- one broad class of phenomena covering several sub-classes
  morphology, grammatical constraints
- concentrating on a particular phenomenon
  ambiguity of pronouns, nouns, conjunctions

Designed CTSs

Engaging Content
Engaging People

phenomena

- morphology
- grammar
- ambiguity
- discourse
- gender bias
- ....

Designed CTSs

Engaging Content
Engaging People

generating entries

- manually
- automatically
- something in between
- use and adapt some existing corpora

Designed CTSs

Engaging Content
Engaging People

evaluation

- manual
- automatic
- something in between

Large taxonomy CTSs

- English-to-French [Isabelle et al., 2017]
- German↔English [Burchardt et al., 2017]
- created to compare NMT with PBMT (and RBMT)
- top-down decision about the phenomena:

  based on linguistic knowledge about the differences between the source and target languages

English-to-French [Isabelle et al., 2017]

- one of the first modern CTSs
- created to compare NMT with PBMT
- 108 manually crafted entries
- manual evaluation
- publicly available
  `https://www.aclweb.org/anthology/attachments/D17-1263.Attachment.zip`

English-to-French

entries

- English source sentence
- French reference translation
- three French MT outputs
- human annotations ("correct" or "incorrect")

Large taxonomy CTSs
English-to-French

phenomena

- morpho-syntactic divergences
  subject-verb agreement, subjunctive
- lexico-syntactic divergences
  argument switching, noun compounds, idioms,...
- syntactic divergences
  yes-no question syntax, clitic pronouns, stranded prepositions, ...

Large taxonomy CTSs

English-to-French

example of an entry for subject-verb agreement

		correct?
Src	The repeated **calls** from his mother **should have** alerted us.	
Ref	Les **appels** répétés de sa mère aur**aient** dû nous alerter.	
PBMT	Les appels répétés de sa mère aurait dû nous a alertés.	no
NMT	Les appels répétés de sa mère devr**aient** nous avoir alertés.	yes
Google	Les appels répétés de sa mère aur**aient** dû nous alerter.	yes

Large taxonomy CTSs
English-to-French

main findings

- NMT generally better than PBMT
- idiomatic expressions are the main shortcoming of NMT

potentials and limitations

- – manually crafted, not easily scalable
- + available manual annotations can help development of automatic methods

German↔English [Burchardt et al., 2017]

- manually crafted test set
- aims to investigate general MT performance against a wide range of linguistic phenomena
- about 5,000 entries
- semi-automatic evaluation
- not publicly available

Large taxonomy CTSs

German↔English

entries

- source sentence
- broad phenomenon category
- specific phenomenon
- MT output
- post-edited MT output

compiled from various corpora

- parallel corpora
- grammatical resources (TSNLP Grammar Test Suite)
- online lists of typical translation errors

Large taxonomy CTSs
German↔English

phenomena

- 14 broader categories
- 106 fine-grained phenomena
- each phenomenon represented by at least 20 entries

Large taxonomy CTSs
German↔English

broader categories

- ambiguity
- composition
- function words
- long distance dependencies & interrogative
- multi-word expressions
- named entities & terminoology
- subordination
- verb tense/aspect/mood
- verb valency

Large taxonomy CTSs
German↔English

Engaging Content
Engaging People

two examples: MWE and non-verbal agreement

Category	MWE	Non-verbal Agreement
Phenomenon	Idiom	Coreference
Source	Lena **machte sich** früh **vom Acker.**	Lisa hat Lasagne gemacht, **sie** ist schon im Ofen.
MT	Lena [left **the field** early].	Lisa has made lasagne, **[she]** is already in the oven.
PE	Lena [left early].	Lisa has made lasagne, **[it]** is already in the oven.

Large taxonomy CTSs
German↔English

semi-automatic evaluation

- check the match with the reference translation (post-edited MT output)
- if no match, evaluate manually

Large taxonomy CTSs
German↔English

extensions [Macketanz et al., 2018, Avramidis et al., 2019]

- increased focus on verb tenses, aspects and mood
- publicly available example (237 entries)
  `https://github.com/DFKI-NLP/TQ_AutoTest/example`
- semi-automatic evaluation based on a set of hand-crafted regular expressions
  - match correct translations ⇒ correct
  - match incorrect translations ⇒ incorrect
  - no match ⇒ manual inspection
- participated in the WMT-18 and WMT-19 test suite tasks

Morphology

Morpheval test suites [Burlot and Yvon, 2017, Burlot et al., 2018]

- English→{Latvian,Czech,German,Finnish}, Turkish→English
- created for assessing morphological competence of MT systems
- 500 entries for each of the three morphological features
- participated in the WMT-18 test suite task
- fully automatic generation and evaluation (no manual checking)
- publicly available
  https://github.com/franckbrl/morpheval
  https://github.com/franckbrl/morpheval_v2
  https://github.com/Helsinki-NLP/en-fi-testsuite

Morpheval CTSs

morphological features

- contrast:
  morphological variants of the same lemma
  (tense, person, case, gender, polarity, ...)
- agreement consistency:
  different POS (pronouns or nouns+adjectives) in the same context
- lexical consistency:
  noun, verb and adjective lexical variations of the same inflection in the same context

Morpheval CTSs

example of morphological contrast

polarity	I am **hungry**.
	I am **not hungry**.
tense	The thing **horrifies** me.
	The thing **horrified** me.
tense	We **did** it.
	We **will do** it.

The translation should reflect the form in the source.

Morpheval CTSs

example of agreement consistency

I see **him**.	PRON
I see a **crazy researcher**.	ADJ N
I see a **good friend**.	ADJ N
I see a **happy linguist**.	ADJ N

All pronouns, adjectives and nouns should be translated into the same case in the target language.

Morpheval CTSs

example of lexical consistency

I agree with the **president**.
I agree with the **director**.
I agree with the **minister**.
I agree with the **driver**.
I agree with the **painter**.

All nouns should be translated into the same case in the target language.

Morpheval CTSs

Engaging Content
Engaging People

entries

- short source sentences from WMT News data
- reference (base) translation
- translation variant(s)

Morpheval CTSs

Engaging Content
Engaging People

generation

- collect a large number of short source and target language sentences (length<15) containing a source feature of interest
- generate translation variant(s)
- compute an average language model (LM) score for each entry (base, variants)
- remove the 33% entries with the lowest LM scores
- randomly select 500 entries for the final test set

Morpheval CTSs

evaluation – contrast

- the differences between the base and the variant translations encode the examined contrast
  ⇒ correct entry
- the base and the variant translations are identical or their differences are irrelevant to the examined contrast
  ⇒ incorrect entry
- resulting score: accuracy averaged over all entries

Morpheval CTSs

evaluation – consistency

- all variants have the same morphological features
  = highest possible consistency of an entry
  ⇒ entropy = 0
- each variant contains a different morphological feature
  = lowest possible consistency of an entry
  ⇒ entropy = 1
- resulting score: average entropy over all entries

Morpheval CTSs

Engaging Content
Engaging People

main findings

- generally, systems with high global quality show a good morphological competence
- rule-based systems have much higher morphological competence than overall quality
- consistency features not correlated with human judgments of overall quality

Morpheval CTSs

potentials and limitations

- generation and evaluation are fully automatic
  - + fast, ability to generate large test sets
  - – relies on automatic morphological analysers in target language and heuristics
    → prone to errors
- resulting scores are based on average accuracy/entropy
  - – it is possible to do more refined analysis (e.g. frequent vs. rare words, etc.)

Grammar

- English-to-German
  how grammatical is NMT output [Sennrich, 2017]
- English-to-Czech
  grammatical constraints [Cinková and Bojar, 2018]

Grammar

English-to-German NMT [Sennrich, 2017]

- assessing five aspects of German grammar
- created for comparing BPE- and character-based NMT
- 97,000 entries
- automatic generation and evaluation
- publicly available
  `https://github.com/rsennrich/lingeval97`

English-to-German NMT

entries

- compiled from WMT News data
- source sentence
- reference (correct) translation
- contrastive (incorrect) translation

English-to-German NMT

generation

- find source and target pairs with phenomena of interest
- for each, generate a contrastive target sentence with one artificially generated error

English-to-German NMT

Engaging Content
Engaging People

phenomena

- noun-phrase agreement
- subject-verb agreement
- separable verb particle
- polarity
- transliteration

Grammar

English-to-German NMT

examples of each grammatical phenomenon

	English	German correct	German contrastive
NP agr.	**of the** American **congress**	**des** amerikanischen **Kongresses**	**der** amerikanischen **Kongresses**
SV agr.	that the **plan will** be approved	dass der **Plan** verabschiedet **wird**	dass der **Plan** verabschiedet **werden**
V particle	he is **resting**	er **ruht** sich **aus**	er **ruht** sich **an**
polarity	the timing is **uncertain**	das Timing ist **unsicher**	das Timing ist **sicher**
translit.	Mr. **Ensign's** office	Senator **Ensigns** Büro	Senator **Enisgns** Büro

English-to-German NMT

contrastive automatic evaluation

based on the ability of NMT systems to assign scores to arbitrary sentence pairs

- the system assigns higher probability to the correct reference translation than to the contrastive translation
  $\Rightarrow$ correct entry
- the other way round:
  $\Rightarrow$ incorrect entry
- resulting score: accuracy

English-to-German NMT

main findings

- character-based decoders are better for processing unknown names
- BPE is better for morpho-syntactic agreement over long distances
- the most challenging error type is the omission of negation markers

Grammar

English-Czech constrasts [Cinková and Bojar, 2018]

- designed for comparing English-to-Czech MT systems
- participated in the WMT-18 test suite task
- 3235 entries
- automatic generation and evaluation
- available via LINDAT-CLARIN repository
  `http://hdl.handle.net/11234/1-2856`

English-Czech contrasts

entries

- automatically selected from Prague Czech-English Dependency Treebank 2.0 (PCEDT)
- English source trees
- Czech reference trees

English-Czech contrasts

phenomena

- verb related
- two types of English clauses which can have many variations in Czech:
  - English gerundial clause (and other -ing forms)
  - English infinitive clause

English-Czech contrasts

examples: gerundial clause

en	He was surprised by the reaction, **calling** it frenetic.
cs	He was surprised by the reaction, **and he called** it frenetic.
en	Consider **adopting** your spouse's name.
cs	Consider **the adopting** of your spouse's name.

example: infinitive clause

en	she agreed **to strike** him
cs	she agreed **with the striking** of him
cs	she agreed **with it that she would strike** him

English-Czech contrasts

automatic evaluation

find the Czech translation of the English part related to the phenomenon
(using automatic word alignments between the English surface tree and Czech surface and deep trees)

- Czech morpho-syntactic properties match the reference
  ⇒ correct entry
- Czech morpho-syntactic properties do not match the reference
  ⇒ incorrect entry
- Czech expression not found
  (e.g. due to alignment problems)
  ⇒ unknown entry

English-Czech contrasts

potentials and limitations

- no manual annotations
- relying on reference translations,
  which could lead to overly pessimistic results
  ("bad" entries are not necessarily unacceptable)

Ambiguity

- pronouns (discourse-related)
  [Guillou and Hardmeier, 2016, Guillou et al., 2018, Müller et al., 2018]
- nouns (lexical ambiguity)
  [Rios Gonzales et al., 2017, Rios Gonzales et al., 2018, Raganato et al., 2019]
- conjunctions (related to the sentence structure)
  [Popović and Castilho, 2019, Popović, 2019]

Ambiguous pronouns

Engaging Content
Engaging People

PROTEST test suite [Guillou and Hardmeier, 2016]

- English-to-French
- 250 manually crafted entries
- semi-automatic evaluation
- publicly available
  `https://bitbucket.org/lianeg/protest/src/master/`

Ambiguous pronouns

PROTEST

entries

- English transcriptions of TED Talks
- their French reference translations
- context for each sentence is available on the document level

Ambiguous pronouns
PROTEST

types of pronouns

- anaphoric (it/they)
  I have a bicycle. **It** is red.
- event (it)
  X invaded Y. **It** resulted in war.
- pleonastic
  It is raining.
- addressee reference (you/your(s))
  - individual person, formal or informal
  - group of people
  - people in general

Ambiguous pronouns
PROTEST

evaluation

- automatic:
  if the target pronoun is in the reference translation
  ⇒ correct
- if not
  ⇒ manual evaluation

Ambiguous pronouns
PROTEST

extensions

- English-to-German
- TED talks and News
- 200 entries
- focus on "it" and "they"
- participated in the WMT-18 test suite task
- main finding:
  inter-sentential context is the most challenging

Ambiguous pronouns

Engaging Content
Engaging People

inter-sentential [Müller et al., 2018]

- English-to-German NMT
- pronoun "it"
- 12, 000 entries
- automatic generation and evaluation
- intended for assessment of context-aware NMT systems
- publicly available
  `https://github.com/ZurichNLP/ContraPro`

Ambiguous pronouns

inter-sentential

entries

- automatically extracted from OpenSubtitles
- English source text
- German reference (correct) translation
- German contrastive (incorrect) translation
  - contains incorrect pronoun
  - grammatically correct in isolation (without the context)

Ambiguous pronouns

inter-sentential

generation

- three possible translations of English "it" into German: "er" ("he"), "sie" ("she") or "es" ("it")
- based on POS tags and automatic word alignments
- contrastive translations:
  change the pronoun in the German reference

Ambiguous pronouns
inter-sentential

contrastive automatic evaluation

same as [Sennrich, 2017]

- the system assigns higher probability to the correct reference translation than to the contrastive translation
  ⇒ correct entry
- the other way round:
  ⇒ incorrect entry
- resulting score: accuracy

Ambiguous pronouns

inter-sentential

main findings

- BLEU scores do not award context-aware models
- these models have much better accuracy on the designed pronoun test suite
- the best model yields +16% accuracy

Lexical ambiguity (nouns)

- German-to-{English,French} NMT
  [Rios Gonzales et al., 2017]
- German-to-English, model-agnostic
  [Rios Gonzales et al., 2018]
- English↔{German,Finnish,Russian,Lithuanian},
  English-to-Czech
  both NMT and model-agnostic
  [Raganato et al., 2019]

Lexical ambiguity (nouns)

German-to-{English,French} NMT [Rios Gonzales et al., 2017]

- 7200 entries for German-to-English
- 6700 entries for German-to-French
- created to assess word sense embeddings for NMT
- semi-automatic generation
- automatic contrastive evaluation [Sennrich, 2017, Müller et al., 2018]
- publicly available `https://github.com/a-rios/ContraWSD`

German-to-{English,French} NMT

entries

- extracted from WMT News, OpenSubtitles, GlobalVoices, EU Bookshop, MultiUN parallel corpora
- source segments
- reference translations
- one or more contrastive examples
  - German-to-English: 3.5 per entry
  - German-to-French: 2.2 per entry

Lexical ambiguity (nouns)
German-to-{English,French} NMT

Engaging Content
Engaging People

generation

- automatically extract ambiguous word pairs from PBMT phrase tables
- clean the lists manually
- use the lists to automatically extract sentence pairs from parallel corpora
- automatically generate contrastive examples

German-to-{English,French} NMT

example of German-to-English entry

source	Ich stellte mich in die **Schlange** für Ausländer.
reference	I got in the **line** for foreigners.
contrastive1	I got in the **snake** for foreigners.
contrastive2	I got in the **serpent** for foreigners.

German-to-{English,French} NMT

main findings

- rarely seen words are challenging (accuracy less than 50%)
- for many words, a larger context is necessary for disambiguation

German-to-English

modifications [Rios Gonzales et al., 2018]

- evaluation process modified to be model-agnostic
- 3249 entries
- participated in the WMT 18 test suite task

German-to-English

Engaging Content
Engaging People

evaluation

- only instances of the correct translations are found
  ⇒ correct
- only instances of the incorrect translations are found
  ⇒ incorrect
- both correct and incorrect translations are found
  ⇒ manual inspection
- none of the two translations is found
  ⇒ manual inspection

Manual inspection required for about 5% of entries.

Lexical ambiguity

German-to-English

main findings

16 German-to-English WMT-18 systems
+ top ranked WMT-16 and WMT-17 systems

- accuracy varies from 43% to 94%
- unsupervised systems are at a clear disadvantage
- translation models have improved since 2016 and 2017

Lexical ambiguity

MuCoW test suite [Raganato et al., 2019]

- 16 language pairs
- about 250, 000 entries
- language-independent automatic generation
- two types of automatic evaluation
- participated in the WMT-19 test suite task
- publicly available
  `https://github.com/Helsinki-NLP/MuCoW`

Lexical ambiguity

MuCoW test suite

two sets for two evaluation methods

- "scoring test suite"
  - 11 language pairs, 240, 000 entries
  - contrastive automatic evaluation
  - suitable only for NMT
- "translation test suite"
  - 9 language pairs, 15, 600 entries
  - automatic check for correct and incorrect words based on word lists
    - only automatic check
    - no manual check for unclear cases
  - model-agnostic

MuCoW test suite

Engaging Content
Engaging People

generation

- identify list of words and their translations from OPUS parallel corpora using automatic word alignments
- use BabelNet (a wide-coverage multilingual encyclopedic dictionary) to cluster word senses
- refine clusters with sense embeddings

MuCoW test suite

Engaging Content
Engaging People

main findings

- the systems yield high precision for in-domain translations
- especially when translating from English into a morphologically rich language
- out-of-domain disambiguation is still challenging

Ambiguous conjunctions

- source languages:
  English, French, Portuguese
- not bounded to any specific target language
  can be used for any target language with the given conjunction ambiguity
- participated in WMT 2019 test suite task
  {English,French}-to-German
- semi-automatic creation
- semi-automatic evaluation
- publicly available
  `https://github.com/m-popovic/evaluating-ambiguous-conjunctions-MT`

Ambiguous conjunctions

entries

- short source English, French and Portuguese sentences (up to 20 words)
- expected target conjunction in German, Spanish, Serbian and Croatian
  (so far tested on these target languages)
- no reference translations available

Ambiguous conjunctions

phenomenon

- English conjunctions "but" and "and" (and their French and Portuguese equivalents) can be translated in two different variants in certain target languages
- depends mainly on the sentence structure
- about 1000 entries for "but"
- about 250 entries for "and"

Ambiguous conjunctions

ambiguity of "but"

in German, Spanish, Serbian, Croatian:

- the first (more frequent) variant can be used after either a positive or a negative clause
- the second variant is used after a negative clause when expressing a contradiction

#1 (aber, pero, ali)	#2 (sondern, sino, nego/već)
We wanted to go to the beach, **but** we went back to the hotel.	We didn't want to go to the hotel, **but** to the beach.

Ambiguous conjunctions

ambiguity of "and"

in Serbian and Croatian:

- the first (more frequent) variant is used to connect non-contrasting actions or ideas
- the second variant is used to indicate that the two connected facts are different

#1 (i)	#2 (a)
The walls **and** the door are white.	The walls are white **and** the door is black.

Ambiguous conjunctions

automatic evaluation

- only the correct conjunction is found
  ⇒ correct
- only the opposite conjunction is found
  ⇒ incorrect
- both conjunctions are found
  ⇒ manual inspection
- none of the two conjunctions are found
  ⇒ manual inspection

manual inspection (for about 2% of entries)

the structure of a sentence is correct
⇒ the sentence is correct

Ambiguous conjunctions

main findings

- "and" is more difficult to disambiguate than "but"
- the first (more frequent) variants are generally not problematic
- most frequent error for the second variant = substitution by the first variant
- related to the amount of the training data

Discourse

- English-to-French [Bawden et al., 2018]
- English-to-Croatian [Šoštarić et al., 2018]
- English-to-Czech [Bojar et al., 2019]

Discourse

English-to-French NMT [Bawden et al., 2018]

- two manually crafted test sets
- coreference and cohesion/coherence
- automatic contrastive evaluation
- aim to assess the integration of linguistic context in NMT
- publicly available `https://diamt.limsi.fr/eval.html`

Discourse

English-to-French NMT

entries

- manually compiled from OpenSubtitles
- contains source text with reference and contrastive translations
- each entry necessarily needs the previous context (in the source and/or the target language) for disambiguation
- 200 entries in each of the two test sets

English-to-French NMT

coreference test set

- aims at evaluating integration of the target side context
- each entry is defined by
  - a source sentence with an anaphoric pronoun "it" or "they"
  - its preceding context containing the pronoun's nominal antecendent
- target pronouns are evenly distributed across number and gender

English-to-French NMT

coherence/cohesion test set

- each entry is defined by
  - a source sentence with one ambiguous word
  - its previous context
- the context might be found on the source side, the target side, or both

English-to-French NMT

main findings

- separate encoders for current and previous sentence are not the best option
- the best strategy:
    - use the previous source sentence as an auxiliary input
    - decode both the current and previous sentence

Discourse

English-to-Croatian [Šoštarić et al., 2018]

- four phenomena focused on unaligned pronouns and determiners
- bottom-up decision about the phenomena:
  1. analysis of both human translations and MT outputs
  2. four problematic divergent patterns/phenomena found
  3. test suites constructed for these phenomena
- automatic extraction with manual checking
- manual evaluation (error analysis)
- available via LINDAT-CLARIN repository `http://hdl.handle.net/11234/1-2855`

English-to-Croatian

entries

- compiled from DGT, SETimes news and TED talks
- 1899 sentence pairs in total
- each marked with the phenomenon tag

English-to-Croatian

phenomena

- Croatian relative pronoun not present in English

en	a resealable bag
hr	vrećica **koja** se može ponovno zatvoriti
en-gloss	bag **which** REFL can again to-seal

- Croatian alternatives for English definite articles (demonstratives or possessives)

en	to address **the** problem, ...
hr	kako bi se nosio s **ovim** problemom
en-gloss	in-order-to would REFL deal with **this** problem

English-to-Croatian

phenomena

- English "it" as subject of a passive or expletive

en	**it** is necessary to make them
hr	potrebno ih je stvoriti
en-gloss	necessary them is to-make

- English possessive pronouns without Croatian equivalent

en	Shortly after **their** arrival, the royal couple...
hr	Nedugo nakon dolaska, kraljevski par...
en-gloss	Shortly after arrival, royal couple...

English-to-Croatian

main findings

comparing three MT systems trained on publicly available data (one PBMT and two NMT)

- all systems perform unsatisfactory on the examined phenomena
- NMT system without BPE performs better than the BPE-based one

Discourse

English-to-Czech [Bojar et al., 2019]

- aims at evaluating document-level language phenomena
- 101 documents with cross-sentence discourse relations
- bottom-up decision about phenomena
  (learning from actual errors in MT outputs)
- manual creation and evaluation
- participated in WMT-19 test suite task
- not yet publicly available

English-to-Czech

coherence phenomena

- topic focus and word order
- position of discourse connectives
- alternative lexicalisations of (multi-word) discourse connectives

English-to-Czech

Engaging Content
Engaging People

main findings

- five systems from WMT-19 evaluated
- no systematic differences observed

Discourse

English-to-Russian NMT [Voita et al., 2019]

- focuses on correct context-agnostic translations which are incorrect in the given context
- bottom-up decision about the phenomena:
  1. analysis of the natural test sets
  2. three context-related problematic patterns/phenomena were found
  3. test suites for these phenomena were created
- semi-automatic creation
- automatic contrastive evaluation
- publicly available: `https://github.com/lena-voita/good-translation-wrong-in-context/tree/master/consistency_testsets`

English-to-Russian NMT

Engaging Content
Engaging People

phenomena

- politeness
  (formal or informal "you")
- ellipsis
  (noun phrase inflection and verb sense)
- lexical cohesion
  (transcription of named entities)

English-to-Russian NMT

entries

- compiled from OpenSubtitles
- 300 entries for politeness of "you"
- 500 entries for noun phrase inflections with ellipsis
- 500 entries for verb sense with ellipsis
- 2000 entries for lexical cohesion
- each entry contains one or more contrastive examples

- contrastive examples:
  correct at the isolated sentence level,
  but not in the given inter-sentential context

English-to-Russian NMT

example of politeness

We haven't really spoken much since **your** return.
Tell me, what's on **your** mind these days?

- both possessive pronouns "your" as well as the verb "tell" should be translated into the same politeness variant (either second singular or second plural)
- contrastive examples:
  mixtures of second singular and second plural

English-to-Russian NMT

example of ellipsis and noun phrase inflection

You call her your friend but have you been to **her home**? **Her work**?

- both "her home" and "her work" have to be translated into the same case determined by the first sentence
- contrastive examples:
  "Her work" in isolation can be translated into any case

English-to-Russian NMT

example of ellipsis and verb sense

Veronica, thank you, but you **saw** what happened.
We all **did**.

- second sentence requires past tense of the verb "see" in the Russian translation
- contrastive examples:
  second sentence with other verbs in the past tense

English-to-Russian NMT

example of lexical cohesion

Go check, **Fran**, get up.
Fran, it is coming to the living room!

- the Russian transcription of "Fran" should be same in both sentences
- contrastive examples:
  mixtures of correct transcriptions

English-to-Russian NMT

Engaging Content
Engaging People

main findings

- several context-aware models developed and compared to a context-agnostic baseline
- context-aware models achieve better accuracies on the CTSs
- indistinguishable BLEU scores

Gender bias

evaluating gender of nouns [Stanovsky et al., 2019]

- does MT rely on gender stereotypes or on the given (intra-sentential) context?
- 3888 English sentences designed to test gender bias in coreference resolution
  - Winogender `https://github.com/rudinger/winogender-schemas`
  - WinoBias `https://github.com/uclanlp/corefBias/tree/master/WinoBias/wino` datasets
- no reference translations in any language
- automatic evaluation
- publicly available `https://github.com/gabrielStanovsky/mt_gender`

Gender bias

Engaging Content
Engaging People

examples of two entries

The **doctor** asked the nurse to help **her** in the procedure.
The **doctor** asked the nurse to help **him** in the procedure.

The lawyer yelled at the **hairdresser** because **she** did a bad job.
The lawyer yelled at the **hairdresser** because **he** did a bad job.

- stereotypical and non-stereotypical variants
- the information about the gender of the noun is given by the pronoun (her, him) in the very same sentence

Gender bias

automatic evaluation

- translate the test set into a target language with grammatical gender
- automatically align source and MT output using the fast_align tool
- identify gender in the target language using off-the-shelf morphological analysers or simple heuristics
- output:
  accuracy

Gender bias

main findings

- tested on eight target languages with grammatical gender (es, fr, it, ru, uk, ar, he, de)
- all 6 tested MT models are significantly prone to rely more on gender stereotypes than on the given context

limitations stated by the authors

- artificially created data set
- medium size (can lead to overfitting if used for training)
- future work:
  more entries, looking for natural samples

Summary

- a number of different CTSs has been developed
- covering a number of distinct or related phenomena
- covering a number of (European) language pairs
- designed/used for different applications

Applications

- comparing distinct MT systems
  (e.g. WMT test suite task, PBMT vs. NMT)
- testing a particular competence of MT systems in general
  (e.g. gender bias, ambiguous conjunctions)
- assessing targeted modifications of an NMT system
  (e.g. context-aware NMT models)

Summary

two types of modern CTSs

- model-agnostic CTSs
  can be used for any MT architecture
  (even for human translations)
- contrastive CTSs
  specifically designed for NMT
  ← first CTSs were specifically designed for RBMT

Testing NMT systems

- testing NMT outputs:
  model-agnostic or contrastive CTSs

- testing NMT components:
  other type of evaluation, but also requires appropriate test sets

Testing NMT components

looking into a neural network

- neural networks have a complex architecture
- how a particular part of neural network captures a particular (linguistic) phenomenon?

How to test a network component on a phenomenon?

preparation

- define phenomenon of interest
- define an appropriate classification task for the given phenomenon
- define a classifier for this task
  (a simple feed-forward neural network is often used, but any type of classifier is possible)

How to test a network component on a phenomenon?

classification data-set

- prepare a labelled data set appropriate for the classification task
- separate in into training and validation set

How to test a network component on a phenomenon?

Engaging Content
Engaging People

training the classifier

- give the training part of the classification data set to the NMT system as input
- extract the output of the NMT component of interest
- train the classifier on this output

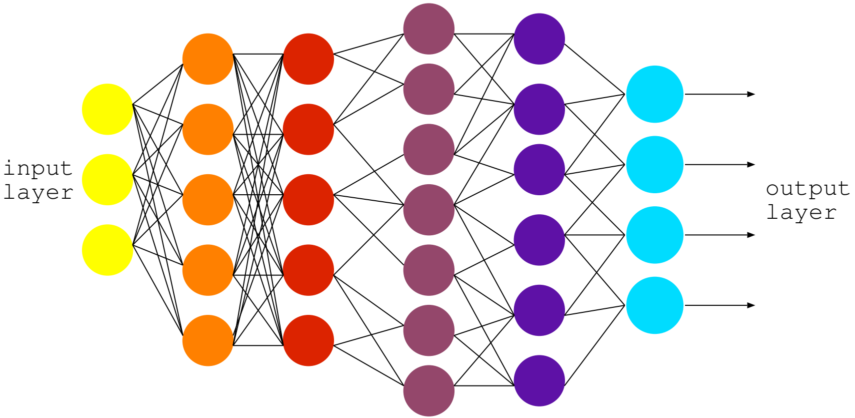
How to test a network component on a phenomenon?

(finally) testing

evaluating the output of the classifier

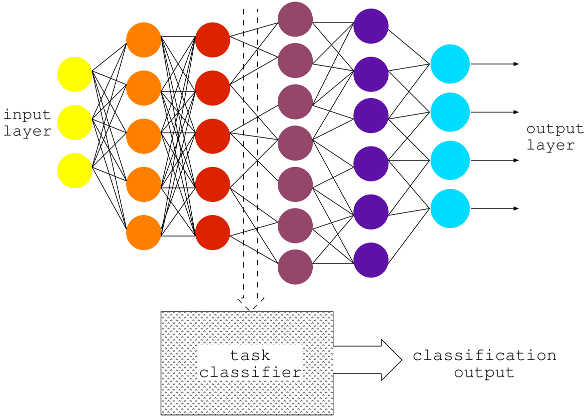
- the higher accuracy ⇒ the better capability of the component to capture the phenomenon

A very simplified neural network



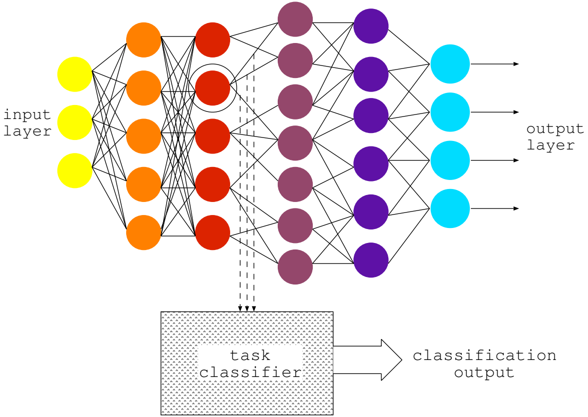
input layer = source
output layer = translation

Testing the red layer



classifier input = output of the red layer

Testing the second node in the red layer



classifier input = output of the second node in the red layer

Data-sets for the classifier

- the classification task should represent well the phenomenon of interest
- the classification data-set should be appropriate for this task
- it has to be big enough to be split into training and validation part

Data-sets for the classifier

Engaging Content
Engaging People

relation to CTSs

- classification data-set is not necessarily a test suite
  - so far, no results on test suites reported in the literature
  - only standard test sets with morphological, syntactic or semantic tags have been used
- test suites can be used if sufficiently large
  - the larger the better
  - 100 or 500 entries will not work well

Tested phenomena

morphology [Belinkov et al., 2017a, Durrani et al., 2019]

- task: full morphological POS tagging (VERB-Sing-2nd-Past, NOUN-Plur-Genitive, ...)
- reasoning:
  - the tested word representations are able to distinguish full morphological POS tags

  ⇒ they are well capable of capturing the word morphology

Tested phenomena

syntax [Durrani et al., 2019]

- task: syntactic tagging (VP, NP, ...)

semantics [Belinkov et al., 2017b, Durrani et al., 2019]

- task: semantic tagging (roles, events, quantifiers, ...)

The same reasoning as for the morphology.

Main findings

- lower layers of the encoder better capture word structure
- decoder learns very little about word structure
- higher encoder layers better capture semantics
- character-based representations are better than BPE

Engaging Content
Engaging People

3. Practical Aspects

Practical aspects

Engaging Content
Engaging People

Thinking about creating your own CTS?

Practical aspects

Thinking about creating your own CTS?

we cannot
tell you what to do exactly

we can
explain what you should think about

Practical aspects

Thinking about creating a new CTS?

- go ahead, it is not very difficult!
- not very easy, either:

  many aspects to think about in advance

I BET SHE IS THINKING OF OTHER MEN...
SO MANY PHENOMENA TO EVALUATE WITH TEST SUITES, HOW CAN I DECIDE?
imgflip.com

Practical aspects

Thinking about participating in the WMT test suite task?

on top of general aspects:

- check translation directions for the given year
- ensure a fast evaluation process

  !! 5–20 MT outputs per translation direction

message from the organisers at WMT-19:

- human parity

⇒ include human translations in the evaluation

our addition: include details about these human translations

Things to think about

- What to evaluate at all?
- How many phenomena?
- Structure of entries?
- How to create entries?
- Model agnostic or designed for specific architecture?
- Evaluation procedure?

What to evaluate?

the decision about which phenomena are of interest is far from trivial

- what is important for the system development?
- what is important for the application of the system?
- what is (potentially) problematic for the given translation direction?

What to evaluate?

Engaging Content
Engaging People

two main strategies

- top-down:
  - start from theoretically problematic linguistic concepts
  - start from concepts important for the task at hand
- bottom-up:
  - look for problematic patterns in standard test sets

What to evaluate?

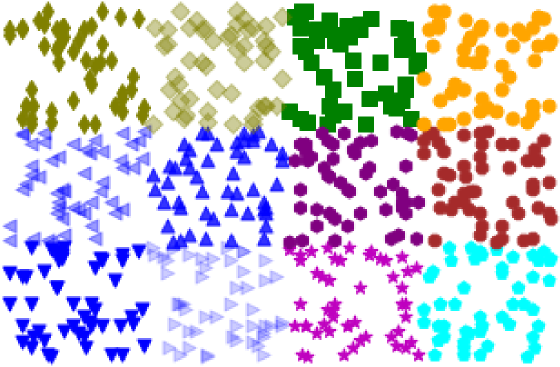
what kind of phenomena is appropriate?

- whatever seems important
- the phenomena are usually linguistically motivated, but it is not necessary
- also depends on the exact definition of "linguistic"
  - is punctuation a linguistic phenomenon?
  - are out-of-vocabulary words a linguistic phenomenon?

Granularity?

Engaging Content
Engaging People

- many distinct phenomena and their variations



Practical aspects

Granularity?

Engaging Content
Engaging People

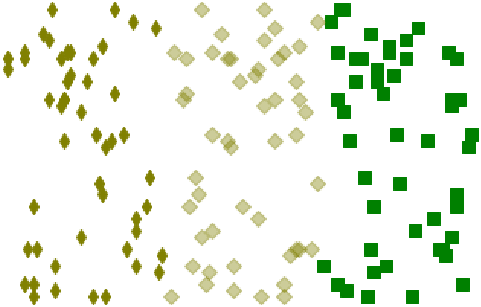
- like

many distinct sweets and their variations!

Granularity?

one broad phenomenon and its variations

- quadrilaterals



Granularity?

one broad phenomenon and its variations

- macaroons

Granularity?

Engaging Content
Engaging People

one particular variation of a single phenomenon

- rhombs

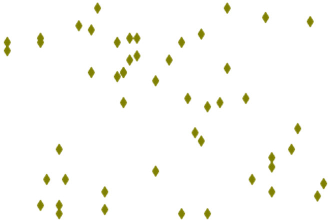
Granularity?

one particular variation of a single phenomenon

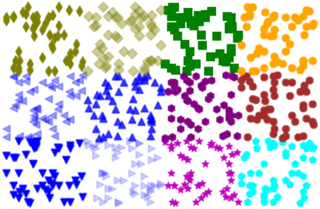
- pistachio macaroons

Practical aspects

Granularity?



might be the first step towards



Granularity?

might be the first step towards

Structure of entries

shorter or longer segments?

- shorter segments have better focus on many phenomena
- less side effects possible in shorter segments
- some phenomena require longer segments
  (long range dependencies)
- longer segments might be more natural

Structure of entries

Engaging Content
Engaging People

allow multiple instances of the phenomenon in one entry?

– this could lead to side effects

+ but it can be interesting

- create both types of entries (single and multiple instances)
- mark each entry with the number of instances
- if possible:
  each multiple instance entry = extension of a single instance entry

Structure of entries

include reference translations?

- if they are available, include them
- if not:
  - if they are really necessary, provide them
  - if not, it's fine

- post-edited MT output or normal human translation?
  - if not comparing the MT system of the PE with other systems, both are fine

Structure of entries

include contrastive examples?

- definitely, if the contrastive NMT evaluation is planned
- possibly, depending on:
  - definition of contrastive examples
  - evaluation procedure

Contrastive examples

contrastive examples are

1. translations with (usually artificially added) errors
2. entries without the phenomenon of interest

they are definitely necessary

1. if the evaluation relies on higher vs lower scores or on a binary classifier ("correct" vs "incorrect")
2. if the phenomenon can be spuriously added to the translation (for example, inserted negation marker)

How to create entries?

Engaging Content
Engaging People

Where to find the desired phenomena?

- extract natural data samples which contain desired phenomena and use them directly as they are
    - parallel multilingual data
    - courses, grammar books (probably monolingual)
- edit and paraphrase natural samples
- completely invent “artificial" entries
- a mixture of everything

How to create entries?

Manually or automatically?

- manually
  - + better control
  - – time and resource consuming
- automatically
  - + fast and easily scalable
  - – more prone to errors
- semi-automatically
  - (try to) take the best of the two worlds
  - ideally:
    automatically with some manual checking/intervention

Model agnostic or for specific architecture?

What is the main purpose of the CTS?

1 comparing distinct systems (e.g. WMT)

1 general purposes

1 no access to the system

⇒ model agnostic

2 measuring progress of own system

⇒ model agnostic works, too

- a tailored test set might be easier to create and/or more appropriate

Evaluation procedure?

- clear evaluation method should be enabled by the structure of entries
  (think about it from the very beginning)
- evaluate only the phenomenon

  — all other factors and errors should be ignored
- afterwards, if possible, compare the results with other types of evaluation
  - overall human or automatic scores
  - error classes
  - ...

Evaluation procedure?

how useful are reference translations?

+ enable automatic evaluation to certain extent
  - the exact extent largely depends on the phenomenon

– can be overly pessimistic
  - always good to have manual inspection for entries with no reference match

+ enable comparison with the overall automatic scores

Evaluation procedure?

Engaging Content
Engaging People

Manually or automatically?

- manually
  - + better control
  - – time and resource consuming
- automatically
  - + fast and easily scalable
  - – more prone to errors
- semi-automatically
  - (try to) take the best of the two worlds
  - ideally:
    automatically with some manual checking/intervention

All in all

Challenge Test Sets are super cool!

- allow for evaluation of specific phenomena
- a number of different phenomena can be tested
- allow for a better understanding of (the) MT system(s)

So go out there and...

Happy Sweet Testing!

Engaging Content
Engaging People

List of publicly available Challenging Test Sets (Test Suites)

Maja Popović and Sheila Castilho

August 2019

- Large Taxonomy
  - English-to-French challenge test set [Isabelle et al., 2017]
    `https://www.aclweb.org/anthology/attachments/D17-1263.Attachment.zip`

- Morphology
  - Morpheval test suites [Burlot and Yvon, 2017, Burlot et al., 2018]
    `https://github.com/franckbrl/morpheval`
    `https://github.com/franckbrl/morpheval_v2`
    `https://github.com/Helsinki-NLP/en-fi-testsuite`

- Grammar
  - English-to-German, five German grammar aspects [Sennrich, 2017]
    `https://github.com/rsennrich/lingeval97`
  - English-to-Czech, two verb-related clause contrasts [Cinková and Bojar, 2018]
    LINDAT-CLARIN repository: `http://hdl.handle.net/11234/1-2856`

- Ambiguity
  - pronouns
    - English-to-{French,German} PROTEST test suite [Guillou and Hardmeier, 2016]
      `https://bitbucket.org/lianeg/protest/src/master/`
    - English-to-German inter-sentential context [Müller et al., 2018]
      `https://github.com/ZurichNLP/ContraPro`

    - nouns
        - German-to-{English,French} [Rios Gonzales et al., 2017, Rios Gonzales et al., 2018]
          `https://github.com/a-rios/ContraWSD`
        - MuCoW test suite [Raganato et al., 2019]
          `https://github.com/Helsinki-NLP/MuCoW`
    - conjunctions ”but” and ”and”
      `https://github.com/m-popovic/evaluating-ambiguous-conjunctions-MT`

- Discourse
    - English-to-French [Bawden et al., 2018]
      `https://diamt.limsi.fr/eval.html`
    - English-to-Croatian [Šoštarić et al., 2018]
      LINDAT-CLARIN repository: `http://hdl.handle.net/11234/1-2855`
    - English-to-Russian [Voita et al., 2019]
      `https://github.com/lena-voita/good-translation-wrong-in-context/tree/master/consistency_testsets`

- Gender of nouns [Stanovsky et al., 2019]
  `https://github.com/gabrielStanovsky/mt_gender`

References

[Bawden et al., 2018] Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 1304–1313, New Orleans, Louisiana.

[Burlot et al., 2018] Burlot, F., Scherrer, Y., Ravishankar, V., Bojar, O., Grönroos, S.-A., Koponen, M., Nieminen, T., and Yvon, F. (2018). The WMT’18 Morpheval test suites for English–Czech, English–German, English–Finnish and Turkish–English. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 550–564, Belgium, Brussels.

[Burlot and Yvon, 2017] Burlot, F. and Yvon, F. (2017). Evaluating the morphological competence of Machine Translation Systems. In *Proceedings of the 2nd Conference on Machine Translation (WMT 17)*, pages 43–55, Copenhagen, Denmark.

[Cinková and Bojar, 2018] Cinková, S. and Bojar, O. (2018). Testsuite on Czech–English Grammatical Contrasts. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 565–575, Belgium, Brussels.

[Guillou and Hardmeier, 2016] Guillou, L. and Hardmeier, C. (2016). PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 16)*, Portorož, Slovenia.

[Isabelle et al., 2017] Isabelle, P., Cherry, C., and Foster, G. (2017). A Challenge Set Approach to Evaluating Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 17)*, pages 2486–2496, Copenhagen, Denmark.

[Müller et al., 2018] Müller, M., Rios Gonzales, A., Voita, E., and Sennrich, R. (2018). A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 61–72, Belgium, Brussels.

[Raganato et al., 2019] Raganato, A., Scherrer, Y., and Tiedemann, J. (2019). The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation. In *Proceedings of the 4th Conference on Machine Translation (WMT 19)*, Florence, Italy.

[Rios Gonzales et al., 2017] Rios Gonzales, A., Mascarell, L., and Sennrich, R. (2017). Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the 2nd Conference on Machine Translation (WMT 17)*, pages 11–19, Copenhagen, Denmark.

[Rios Gonzales et al., 2018] Rios Gonzales, A., Müller, M., and Sennrich, R. (2018). The Word Sense Disambiguation Test Suite at WMT18. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 594–602, Belgium, Brussels.

[Sennrich, 2017] Sennrich, R. (2017). How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain.

[Stanovsky et al., 2019] Stanovsky, G., Smith, N., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 19)*, Florence, Italy.

[Voita et al., 2019] Voita, E., Sennrich, R., and Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 19)*, Florence, Italy.

[Šoštarić et al., 2018] Šoštarić, M., Hardmeier, C., and Stymne, S. (2018). Discourse-Related Language Contrasts in English-Croatian Human and Machine Translation. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 36–48, Belgium, Brussels.