Extrinsic Evaluation of Patent MT: Review and Commentary

Douglas W. Oard
University of Maryland
College Park, MD 20742 USA
oard@umd.edu

Noriko Kando
National Institute of Informatics
Tokyo, Japan
kando@nii.ac.jp

ABSTRACT
There has been a long history of work on the application of Machine Translation (MT) to support cross-language information access for patent collections. Much of this work has leveraged fairly traditional information retrieval evaluation designs as a basis for extrinsic (i.e., task-based) evaluation, but other evaluation designs are also possible. This survey reviews the work to date on extrinsic evaluation of patent MT in cross-language information access applications, identifying gaps in the literature, and formulating some open research questions.

Categories and Subject Descriptors
I.2.7 [Computing Methodologies]: Artificial Intelligence – *natural language processing.*

General Terms
Measurement, Performance, Design, Experimentation.

Keywords
Patents, Machine translation.

1. INTRODUCTION
Machine Translation (MT) is one of the defining technologies of this age of globalization, and in recent years machine translation systems have become increasingly capable. This success has resulted from two broad trends: (1) statistical modeling of language use, and (2) evaluation-guided research. To date, the dominant focus of MT evaluation has been on intrinsic evaluation. That, however, is not the only possible approach, and in this paper we focus on an alternative, one that is more closely grounded in the information access tasks that give this workshop on Evaluation of Information Access (EVIA) its name: extrinsic evaluation, and in particular extrinsic evaluation in the context of some information access task.

According to Sparck Jones [14], "*Intrinsic* criteria are those relating to a system's objective, *extrinsic* criteria are those relating to its function, i.e. to its role in realisation to its set-up's purpose. ' (Author's gloss: a set-up is a system together with the setting in which it is to be used). [...] Thus for a translation system, intrinsic criteria could be ones applying to the quality of the translation, and extrinsic criteria those applying to the ease with which post-editors could tweak these, while for the translation set-up as a whole intrinsic criteria could refer to e.g. to the speed of production of the final translation, and extrinsic criteria to the value/acceptability/utility of the final translation for some purpose such as literature scanning."

The National Institute of Informatics (NII) Testbeds and Community for Information Access Research (NTCIR) project has sponsored evaluations of extrinsic evaluation for Machine Translation (MT) of patents since 2004. The purpose of this brief survey is to review that work with an eye towards identifying additional extrinsic evaluation designs that would be practical in the near term, that could yield useful additional insights, and for which the potential benefits are expected to be sufficient to justify the expected costs. The proposed task designs are meant to be illustrative of the available options.

2. TASK MODELING

Intellectual property professionals distinguish between many types of tasks that could involve searching for and understanding translated content. For purposes of evaluation design, we can usefully group these tasks into three broad categories (expressed here using specific examples of languages and sources):

Technology Survey Task: The task is to determine whether an idea that is described in the English abstract of a scientific paper has been patented in Japan (or whether an idea that is described in the Japanese abstract of a scientific paper has been patented in the USA or in Europe).

Invalidation Task: The task is to determine whether each specific claim made in a Japanese patent application has been previously described in a patent in the USA or Europe (or whether each specific claim in a European patent application has been patented in Japan). The invalidation task is one use case for a **Prior Art Search Task**, which is a more general task that might also be performed prior to the creation of a patent application.

Expedited Review Task: The task is to determine whether specific sources for prior art were considered by patent examiners in Japan when awarding a patent so that patent examiners in the USA or Europe can focus their attention on sources uniquely required to be searched by their respective patent laws. In October 2004 the JPO introduced an Advanced Intellectual Property Network (AIPN) in which the results of patent examinations in Japan are made available to patent offices in other countries as a basis for expedited processing of patents in a family for which the initial application was made in Japan [16]. Both the United States Patent and Trademark Office (USPTO) and the European Patent Office (EPO) participate.
Within these tasks, we can identify three potential uses for Machine Translation:

- To support the **automated detection of required information**. "Detection" can be read narrowly (with

each item either detected or not) or broadly (to include estimation of the relative value of each item). An item may be a set of documents (e.g., a patent family), an individual document (e.g., a prior art patent), or a part of a document (e.g., one—possibly partially—invalidating claim found in a prior art patent).

- To support **manual detection of required information**. Because NTCIR is focused on Information Access research (by which is meant the process of gaining access to specific information in large collections that would otherwise be inaccessible), techniques for supporting manual detection of required information are of interest principally for their potential for employment together with automated detection of required information as part of a complete human-machine system for supporting information access.
- To support the **use of required information** once it has been detected. As with manual detection, techniques for supporting the use of required information are of interest in NTCIR principally for their potential for employment together with information access technology to support the full continuum from information access to information use.

3. TYPES OF EXTRINSIC EVALUATION

Extrinsic evaluation often calls to mind evaluation with users in the loop. While user studies certainly do offer one way of conducting extrinsic evaluation, evaluations based on "canned" task models offer different points on an affordability vs. insight continuum. Looking broadly at evaluation frameworks that are used generally to evaluate the information access task that serves as the basis for extrinsic evaluation at NTCIR, three types of approaches are evident:

Comparative Automatic: The task is to compare the ability of several alternative system designs to produce a high-quality intermediate result (e.g., a ranked list) for which we believe that our quality measure is predictive of the user's ability to get good results (accuracy, speed, and/or satisfaction) when using the system. The key requirements are:

- Two or more MT systems whose results are used by one or more IR systems. Examples include: (1) each MT result is used by a different IR system (the usual CLIR design), (2) all MT results are used by one IR system (the NTCIR-8 PatentMT design), or (3) all MT results are available for use by any IR system (the basic idea in the TRECVID design).
- A "canned" scenario that is representative of the conditions in which the component(s) to be tested would be used. Examples include: (1) topics/documents/relevance judgments (for ad hoc ranked retrieval), (2) training judgments/evaluation judgments (for text classification).

This experiment design can be crafted to produce reusable evaluation resources. A widely discussed sequence of user studies suggest that caution should be exercised when extrapolating from component performance to user performance in cases in which users are able to get good results even with mediocre systems [12].

Comparative Interactive: The task is to compare the user's ability to employ alternative interactive system designs to perform some type of task. The key requirements are:

- A pool (e.g., 16) of representative searchers (e.g., intellectual policy experts with the requisite language skills for the task).
- Two sufficiently capable interactive systems for the task to be performed.

The basic approach is to use a latin square design to identify central tendencies in measurable aspects of user performance (e.g., accuracy or speed) when using a system the presence of topic effects, user effects, and topic-user interaction effects (all of which are modeled as noise). This experiment design does not produce reusable evaluation resources. This design has been used in the TREC interactive track and in iCLEF. Experience suggests that only within-site comparisons can reasonably be made, so every site must have access to two systems (although one could be provided as a baseline by the organizers).

Absolute: The task is to characterize, in some absolute sense, the user's ability to employ some system to perform some specific task. The key requirements are:

- A single system,
- A single user (or user team, if user's work together to accomplish the task), and
- A single task.

Absolute evaluation requires that what has not been found be characterized with reasonable accuracy, and for that reason absolute evaluation can be more expensive than comparative evaluation. Absolute evaluation results can be used for comparative purposes, but the reverse is not generally true. Some absolute evaluation designs (e.g., for automated evaluation of text classification) can yield reusable evaluation resources.

4. MEASURES OF EFFECTIVENESS

As usually formulated, the key question to be answered when evaluating the effectiveness of some process for supporting information access is whether the items required to accomplish the task have been found. The ground truth relevance judgments needed to support such an evaluation can be obtained in one of two ways:

- **Found Judgments.** In patent retrieval, the usual way of obtaining found judgments is to know (from documents produced during the patent application and/or examination processes) which documents some qualified person (e.g., an examiner) actually found to be sufficient to accomplish the task. To date, this has invariably been followed by an implicit assumption that those same documents are also necessary to perform the task; the consequence of known documents being necessary and sufficient is that all other patents can be treated as not relevant.

- **Manual Judgments.** In information retrieval evaluation generally, the usual way of creating relevance judgments is to have domain experts (preferably the creators of each topic) judge the relevance of a sample, to project those results to estimate the density of relevant documents in both the retrieved set (for any system) and the collection. When

the density of relevant documents in the collection is low, unequal (e.g., stratified) sampling is typically used to improve the accuracy of the estimates. The widely used pooling method is a limit case of unequal sampling in which one stratum is not sampled and the prevalence in that stratum is assumed to be zero (which is a reasonable assumption for some types of relative comparisons).

These relevance measures can be used to compute measures of effectiveness that give some insight into the degree to which the goals of an information access task have been satisfied. Two broad classes of evaluation measures have been proposed:

- Decision measures that are **computed as if the ground truth relevance judgments are a complete labeling** of all items [11]. For evaluating sets, common examples include precision, recall, and F_1. For evaluating ranked lists, common examples include Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG) and Patent Retrieval Evaluation Score (PRES). Measures in this class can be computed using only positive relevance judgments (by assuming that unjudged items are not relevant), although such an assumption typically results in values that are suitable only for relative comparisons.

- Decision measures that **explicitly model incompleteness in the ground truth relevance judgments** [13]. For evaluating sets, common examples include estimated precision, estimated recall, and estimated F_1. For evaluating ranked lists, common examples include Binary Preference (bpref), Inferred Average Precision (infAP) and Mean Average Precisoon on ranked lists from which unjudged documents have been removed (MAP'). Measures in this class require both positive and negative relevance judgments.

- **Task-specific measures** that reflect costs and benefits of likely outcomes rather than focusing on individual decisions. For example, Trippe and Ruthven have suggested that risk-adjusted outcome measures might differentially affect the importance of recall in (a) prior art searches conducted to help inform the scope of the claims to be made in a patent application, or (b) prior art searches conducted for an invalidation task [15].

5. EXISTING EVALUATION DESIGNS

Four extrinsic evaluations of Patent MT have been completed to date, and a fifth was at the time of this writing planned for NTCIR-10:

The **NTCIR-3 Patent Retrieval task** included an extrinsic evaluation of patent translation that the organizers called Cross-Language Information Retrieval (CLIR) [4]. The key idea was to view the purpose of Machine Translation (MT) as being to support ranked retrieval of existing patents to identify previously awarded patents related to some topic of interest (which was represented by a news story). The specific design of the task was:

- 31 topics were created in Japanese by intellectual property professionals to be representative of a technology survey task and manually translated into English (and Chinese and Korean, although the Chinese and Korean were not used).
- Japanese terms associated with terms used in the English topic were found by two participating teams, both of which used comparable corpus techniques on a parallel English-Japanese test collection (this amounts of bag-of-words translation).
- Each team used their own patent retrieval system to search the patent collection using their own bag-of-words translation as a bag-of-words query.
- Intellectual property professionals, with and without access to both monolingual and CLIR system results, identified as many relevant patents as possible and all other documents were treated as not relevant.
- Mean Average Precision (MAP) was reported as the evaluation measure.

The **NTCIR-4 Patent Retrieval task** included an extrinsic evaluation of patent translation that the organizers again called CLIR [1]. The key idea was to view the purpose of Machine Translation (MT) as being to support ranked retrieval of existing patents to identify previously awarded patents that invalidate some claim in a new patent application. The specific design of the task was:

- The claims section for each of 34 rejected patent applications was obtained from the Japan Patent Office (JPO) and manually translated into English (and Chinese, although the Chinese was not used).
- Japanese terms associated with terms used in the English claim were found by one participating team by using comparable corpus techniques on a parallel English-Japanese test collection (this amounts of bag-of-words translation).
- A patent retrieval system was used to search the patent collection with the bag-of-words translation as a bag-of-words query.
- Each patent (in Japanese) that was cited in the decision document rejecting the application, and each additional invalidating patent that could be found by professional patent searchers, with and without access to both monolingual and CLIR system results, was treated as a relevant document and all other documents were treated as not relevant.
- Mean Average Precision (MAP) was reported as the evaluation measure.

The **NTCIR-7 PatentMT task** included an extrinsic evaluation of patent translation that the organizers called Cross-Language Patent Retrieval (CLPR) [2]. The key idea was to view the purpose of Machine Translation (MT) as being to support ranked retrieval of existing patents to identify previously awarded patents that invalidate some claim in a new patent application. The specific design of the task was:

- The first claim for each of 124 rejected patent applications was obtained from the JPO and manually translated into English.
- This English claim was transited by MT into Japanese by each of 12 participating teams.

- A standard patent retrieval system was used to search the patent collection with the MT-generated Japanese claim as a bag-of-words query.
- Each patent (in Japanese) that was cited in the decision document rejecting the application was treated as a relevant document, and all other documents were treated as not relevant.
- MAP was reported as an evaluation measure.

The **NTCIR-8 PatentMT task** included an extrinsic evaluation of patent translation using the same CLPR design, this time with 91 rather than 124 claims and 6 participating teams [3]. In NTCIR-7, the claims were selected to be relatively easy (monolingual AP between 0.3 and 0.9); in NTCIR-8 the claims were selected to be relatively hard (monolingual AP below 0.4).

The **NTCIR-10 PatentMT task** includes an extrinsic evaluation of patent translation that the organizers call Patent Examination Evaluation (PEE).[1] The key idea is to view the purpose of MT as being to support making a decision on whether to award a new patent based on an understanding of whether some other (existing) patent invalidates the claims of the new patent application. The specific design of the task is:

- Some number of rejected patent applications to the JPO are selected.
- Bilingual volunteers from the Nippon Intellectual Property Translation Association will serve as the assessors.
- For each rejected patent, the assessor is given:
    - The decision document (in Japanese) that identifies specific facts found in some specific prior patent that led (perhaps in part) to the rejection of the patent application.
    - The translated patent (translated by MT from Japanese to English) in which those specific facts were found.
- The assessor is asked to determine (on a graded scale) whether the degree to which those specific facts could have been ascertained from the translated patent.
- A second version of PEE, in which the prior patent is first manually translated by hand from Japanese to Chinese and then by machine from Chinese to English will also be run.

6. OTHER AVAILABLE RESOURCES

Although to the best of our knowledge the following resources have not yet been used for extrinsic evaluation of Patent MT, the following resources clearly have some potential for that purpose:

The **NTCIR-7 and NTCIR-8 Patent Mining task** created hand-verified associations between 1,200 research papers (644 for NTCIR-7 and 976 for NTCIR-8) in the 255,960-document NTCIR-1 and NTCIR-2 scientific papers abstracts collection (which is available in English and Japanese) and Japanese patent applications published between 1993 and 2007 [7,8]. In the Patent Mining task these associations were used to project International Patent Classification (IPC) from a patent to an abstract, but the same associations could be used in an extrinsic evaluation of MT in which an English abstract from a scientific paper is presented as a query to determine whether a patent application has been filed based on the work reported in that paper (but suppression of the reference to the paper would be required). For some of these patents, it might also be possible to automatically identify patents in the same family in English or Chinese using techniques originally developed for the Patent Mining task.

The **CLEF-2010 and CLEF-2011 Intellectual Property lab** (CLEF-IP) produced a test collection that could be (but has not yet been) used for extrinsic evaluation of Patent MT [9,10]. The test collection includes a patent application as a query document, and citations from various sources to awarded patents as relevance judgments. The query document is available in a single language (English, French, or German) and the awarded patents contain two fields (title and claims) in all three languages. For use in CLIR experiments, the title and claims in the query language would need to be suppressed (that is not done at CLEF, and we are not aware of it yet having been done with this collection).

7. FUTURE EVALUATION DESIGNS

In this section we propose three possible future directions for extrinsic evaluation of Patent MT, ordered by increasing cost and complexity.

To date, all extrinsic evaluations of Patent MT have involved comparative evaluation of automated detection, and in particular ranked patent retrieval, and all have used measures that assume that uncited patents are not relevant.[2] There are two reasons to question whether the implicit assumption in the NTCIR-7 and NTCIR-8 extrinsic evaluation of Patent MT that no uncited patents could have served as a basis for invalidation. First, Fujii et al reported that, on average over 34 topics, found judgments (from citations found in the decision report for rejected patent applications) comprise only 30% of the invalidating patents that could be found by experts (38% for patents that were individually sufficient to invalidate; 24% for patents that were together sufficient to invalidate) [1]. Under such conditions, recall computed using found patents would be therefore substantially understated (although we cannot tell from this data alone whether relative comparisons would remain informative). Second, Lupu et al reported that using manual judgments results in substantially different system ordering from found judgments (from citations found in awarded patents) when Mean Average Precision was computed over 12 topics (thus suggesting that relative comparisons may be less informative than would be desirable) [5]. Together, these results suggest that the relative comparisons reported in NTCIR-7 and NTCIR-8 could have been adversely affected to some extent by incompleteness of the relevance judgments. This is important because extrinsic evaluations (at NTCIR-3 and NTCIR-4) explored only comparable corpus techniques and this the record of extrinsic evaluation for full Statistical MT (SMT) systems relies exclusively on NTCIR-7 and NTCIR-8. We therefore suggest that a study be designed in which both positive and negative judgments are available, and in

[1] http://ntcir.nii.ac.jp/PatentMT-2/

[2] The NTCIR-10 PEE evaluation will be the first departure from that pattern.

which measures such as bpref or infAP that can accommodate incomplete relevance judgments be conducted. Such a study would require some manual relevance judgments.

Second, we suggest that additional studies be designed for the Expedited Review task. The NTCIR-10 PEE evaluation, with its focus on absolute evaluation of interactive detection of invalidating claims, can be viewed as a first step in this direction. A first step in this direction would be to convene a discussion between NII, JPO and either USPTO or EPO to develop a task model that simultaneously achieves sufficiently high fidelity to yield results that can help to guide system development and sufficient simplicity to permit affordable evaluation (and, if possible, reusable evaluation resources). A natural second step, building on the NTCIR-10 PEE task, might be to perform comparative evaluation of systems designed to automatically detect the specific types of information required for the Expedited Review task in a large collection of patents from another jurisdiction in the specific national (or regional) context of that jurisdiction. In addition to USPTO or EOP patents, patents from the State Intellectual Property Office (SIPO) of China might also be considered.
A possible third step would be to conduct a comparative interactive evaluation of one or more proposed systems for supporting the Expedited Review task, with the current interactive process serving as a baseline. A comparative interactive evaluation would require significant resources (both in the effort required to configure research systems for comprehensive coverage and in the time required from a substantial number of participants in the study), but Lupu reports that interactive case studies are commonly reported in the professional patent searching literature [6], and some form of interactive evaluation might therefore prove useful as a way of fostering adoption of new technology.

8. CONCLUSION

Although Patent MT has been a focus for extrinsic evaluation for nearly a decade now, much remains to be learned. The field has benefited enormously from the focus on the invalidation task and the use of found judgments that have been the hallmark of patent retrieval evaluation in Japan, Europe and the United States, but as with any evaluation design there are limits to what can be learned using that design. We believe that this is a propitious time to explore new directions. Motivated in part by the creative thinking of the NTCIR-10 PatentMT task organizers with their PEE task, we suggest that focusing on the Expedited Review task could yield both new insights and considerable potential for technology transition. Motivated by the widespread adoption of new evaluation measures for automated detection of required information that can accommodate incomplete relevance judgments, we further suggest that such measures be considered in future extrinsic evaluations of Patent MT. This is an exciting time to be working at the intersection of machine translation, information access, and intellectual property, and we look forward to a second decade of working together that will be as productive as our first decade working together has been.

9. REFERENCES

[1] Atsushi Fujii, Makoto Iwayama and Noriko Kando (2004). Overview of Patent Retrieval Task at NTCIR-4. *Proceedings of NTCIR-4*, Tokyo, June 2-4. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/PATENT/NTCIR4-OV-PATENT-FujiiA.pdf

[2] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto and Takehito Utsuro (2008). Overview of the Patent Translation Task at the NTCIR-7 Workshop, *Proceedings of NTCIR-7 Workshop Meeting*, December 16-19. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C3/PATMT/01-NTCIR7-OV-PATMT-FujiiA.pdf

[3] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya and Sayori Shimohata (2010). Overview of the Patent Translation Task at the NTCIR-8 Workshop, *Proceedings of NTCIR-8 Workshop Meeting*, Tokyo, June 15-18. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/01-NTCIR8-OV-PATMT-FujiiA.pdf

[4] Makoto Iwayama, Atsushi Fujii, Noriko Kando and Akihiko Takano (2003). Overview of Patent Retrieval Task at NTCIR-3. *Proceedings of the Third NTCIR Workshop*, Tokyo, October 8-10. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-OV-PATENT-IwayamaM.pdf

[5] Mihai Lupu, Florina Piroi and Alan Hanbury (2010). Aspects and Analysis of Patent Test Collections, *Proceedings of the Third International Workshop on Patent Information Retrieval*, Toronto, October 26. http://dx.doi.org/10.1145/1871888.1871892

[6] Mihai Lupu (2011). The Status of Retrieval Evaluation in the Patent Domain, in *Proceedings of the 4th Workshop on Patent Information Retrieva*l, Glasgow, October 24. http://dx.doi.org/10.1145/2064975.2064985

[7] Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama and Taiichi Hashimoto (2008). Overview of the Patent Mining Task at the NTCIR-7 Workshop, *Proceedings of the NTCIR-7 Workshop Meeting*, Tokyo, December 16-19. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C3/PATMN/01-NTCIR7-OV-PATMN-NanbaH.pdf

[8] Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama and Taiichi Hashimoto (2010). Overview of the Patent Mining Task at the NTCIR-8 Workshop, *Proceedings of NTCIR-8 Workshop Meeting*, June 15-18. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/01-NTCIR8-OV-PATMN-NanbaH.pdf

[9] Florina Piroi and John Tait (2010). CLEF-IP 2010: Retrieval in the Intellectual Property Domain, CLEF 2010 Labs and Workshops Notebook Papers, Padua, September 22-23. http://clef2010.org/resources/proceedings/clef2010labs_submission_122.pdf

[10] Florina Piroi, Mihai Lupu, Allan Hanbury and Veronika Zenz (2011). CLEF-IP 2011: Retrieval in the Intellectual Property Domain, CLEF 2011 Labs And Workshops Notebook Papers, Amsterdam, September 19-22. http://clef2011.org/resources/proceedings/Overview_CLEF-IP_Clef2011.pdf

[11] Walid Magdy and Gareth J.F. Jones (2010). Examining the Robustness of Evaluation Metrics for Patent Retrieval with Incomplete Relevance Judgments, Proceedings of the 2010

International Conference on Multilingual and Multimodal Information Access, Padua, pp. 82-93.

[12] Catherine L. Smith and Paul B. Kantor (2008). User Adaptation: Good Results from Poor Systems, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore, July 20-24. http://dx.doi.org/10.1145/1390334.1390362

[13] Tetsuya Sekai and Noriko Kando (2008). On Information Retrieval Metrics Designed for Information Retrieval Evaluation with Incomplete Relevance Assessments, *Information Retrieval*, 11(5)447-470.

[14] Karen Sparck Jones (1997). Coherent Approaches to Evaluation, *EAGLES Evaluation Group Workshop*, Brussels, November 26-27. http://www.cst.dk/eagles/workshop/TRECkaren.html

[15] Anthony Trippe and Ian Ruthven (2011). Evaluating Real Patent Retrieval Effectiveness, in *Current Challenges in Patent Information Retrieval*, Springer-Verlag, Berlin. http://dx.doi.org/10.1007/978-3-642-19231-9

[16] Eiichi Yamamoto (2011). Future Plans of Machine Translation System in the JPO (powerpoint slides), Proceedings of the 4th Workshop on Patent Translation, Xiamen, September 23. http://www.mt-archive.info/MTS-2011-Yamamoto.pdf