ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation

www.accurat-project.eu

Project no. 248347

Deliverable D5.3

Report on requirements, implementation and evaluation of usability in application for web authoring

Version No. 1.0

29/06/2012

Document Information

Deliverable number:	D5.3
Deliverable title:	Report on requirements, implementation and evaluation of usability in application for web authoring
Due date of deliverable:	30/06/2012
Actual submission date of deliverable:	29/06/2012
Main Author(s):	Mateja Verlic
Participants:	Tilde
Internal reviewer:	Tilde
Workpackage:	WP5
Workpackage title:	Evaluation of usability in applications
Workpackage leader:	Tilde
Dissemination Level:	**PU**
Version:	V1.0
Keywords:	Translation, suggestions, web authoring, blog

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/ Approval Level**
V0.1	30/04/2012	Draft	Mateja Verlic (ZEM)	Structure of the document	
V0.2	20.6.2012	Working	Mateja Verlic (ZEM)	Introduction, general requirements	
V0.3	28.6.2012	Pre-final	Mateja Verlic (ZEM)	Evaluation results, statistics	

EXECUTIVE SUMMARY

This deliverable describes requirements, implementation and evaluation of usability of Accurat MT in Zemanta's application for web authoring. It also includes evaluation results, which show, that using comparable corpora enhanced machine translation improved the relevance of related articles recommended by Zemanta.

Table of Contents

Abbreviations

Abbreviation	Term/definition
REST	REpresentational State Transfer – software architecture for distributed systems
API	Application Programming Interface (API) – specification of interfaces between software components
SOAP	Simple Object Access Protocol – protocol specification for implementation of Web services
jQuery	Cross-browser JavaScript library
Ajax	Asynchronous JavaScript and XML – web development techniques for creating asynchronous web applications

1. Introduction

Since late 1990s when forums and bulletin boards were replaced by blogging, it has only gained on importance. Blogging is one of the latest forms of web authoring and is more and more merged into our everyday life along with online news portals. A lot of companies write their own blog to attract more customers, people with different expertise want to share their experience and knowledge, artists are publishing their portfolios online and more and more people are blogging for fun and for their personal satisfaction. NielsenMcKinsey company tracked over 181 million blogs around the world (1) and according to the report "State of the blogosphere 2011: Introduction and Methodology" by Technorati (2), 60% of bloggers in blogosphere are hobbyists, blogging for fun, 18% are professional part- and full-timers, 8% of bloggers are corporate bloggers and 13% of bloggers are characterized as entrepreneurs.

Although blogging cycle differ from blogger type to blogger type, a typical blogging cycle starts with an idea, an insight or just a comment on current affairs author wants to share with her readers. Usually bloggers prepare general draft in one of the text editors on their computer, then they copy text, paste it into online editor of their blogging platform, add some images, links, sometimes they add tags denoting the topics of the blog posts and publish it. The part of the cycle, when user pastes the text into editor and publishes it can vary from few minutes to couple of hours – it depends on the quality of blog posts blogger strives to achieve.

Zemanta's role in the blogging process is assisting blogger by recommending content related to the text. This content includes related images, inline links, tags and most importantly related articles, which with a click of the button enable bloggers to blog faster and better. Zemanta suggests articles from other bloggers writing on same or similar topics, so blogger doesn't have to go and search for related articles, instead she can find them next to the blog editor and add them by simply clicking on them. Typically bloggers don't spend too much time in the online editor, what is an important time constraint for e.g. machine translation a web service.

Furthermore, in comparison to other fields or domains where machine translation is needed and wanted, blog publishing happens instantly; readers can access the text only few moments after author hits publish button. This is especially important for services bloggers use when blogging, e.g. searching and adding related posts or images, analyzing text to add related tags. Services have to return at least some (if not all) results in reasonable time, and machine translation web service is not an exception.

In Workpackage 5 Zemanta was responsible for evaluation of developed machine translation methods for use in web authoring application. In following section we will summarize general requirements for web services, especially in web authoring domain, we will report on implementation and finally evaluation of usability in a web authoring application.

The main goal of the evaluation was to find out whether Zemanta recommendation engine returns better results (related articles) for texts using Accurat MT methods than it does for original (not translated) texts.

Evaluation process was organized three parts: in the first part we evaluated translation results for baseline and CC-enhanced MT method for 100 texts for SL-EN, DE-EN and HR-En language pairs. Results have been evaluated using Zemanta's internal evaluation tool Dash. In the second part we analyzed 10 randomly selected files for each language pair and translation method to assess the quality of translation. In third part we implemented WordPress plugin for Zemanta with integrated translation service and included demonstration in a use case.

2. General Requirements

Every web service has to fulfill some basic requirements to be considered useful in the web authoring domain. General requirements here are described from our point of view and in the context of integration and use of machine translation service in Zemanta's widget.

2.1. Machine translation service as a web service

As mentioned before, most of users prepare their texts in desktop text editors and copy/paste their text in online editor to enrich it with images, links and related articles. Because Zemanta widget works from within online editor (e.g. editor in Word Press) , machine translation has to be offered as a web service.

Installing another desktop application to translate text or copy the text in a web form was not an option. In our case author does not have to see the translation results, because translation is only an intermediate step and serves as an input to Zemanta recommendation engine.

2.1.1. REST API

Implementing translation web service as a REST service is recommended, due to its simplicity in comparison to other web service design models such as SOAP or WSDL. Providing an API to translation service it makes possible to integrate the service or use it for batch translations in a client.

2.1.2. Response time

In general machine translation can take ling time to translate text. How long are bloggers willing to wait before they get some results? Average blog posts can take from half an hour to couple of hours – depending on the topic and thoroughness of the author's research. But in any case it is not acceptable. Bloggers are willing to wait for few minutes or as long as it takes them to add images and format the text.

2.1.3. Number of requests limit

Zemanta recommendation engine checks every 10 to 15 seconds if editor contains 300 characters and if it does, it refreshes recommendations (related images, related articles). During these intervals widget can also check whether translation is completed and returns the results.

2.2. Registration

We have to have some information about the identity of the requester of the translation service, especially if translation service will be open to public. Currently translation service requests special token, which is passed as a parameter to the server, but at this point there is no registration page for obtaining the token.

2.3. Availability

Bloggers blog day or night and therefore translation service has to be available 24/7.

If translation service times out, blogger has to be notified and web authoring tool widget should not stop responding.

2.4. Language pairs

Different bloggers writing in different languages might use this service and therefore language pair is sent as a parameter to the service, which then creates translation job with appropriate worker. We evaluated translation service for three language pairs: SL-EN, DE-EN, HR-EN. All three pairs have English as the target language, because Zemanta works with English texts only.

2.5. Amount of text

The length of text bloggers produce vary a lot. It depends on the topic blogger writes about and the purpose of the blog. Some bloggers use minimum text, adding only a line or two of commentary, while others write detailed reviews.

2.5.1. Average blog and average news text

The length of an average blog post or (short) news text is between 200 and 300 words. For texts of these lengths machine translation returns results in a reasonable time (few minutes). Longer texts could cause more problems for the translation service.

2.6. MT quality

The quality of machine translation for web authoring in our case is not about how translated text looks like, but more how much of it has been translated and also how fast it can integrate new concepts and names.

2.6.1. Keywords and named entities

Machine translation method has to be able to translate keywords and named entities. If personal names are not international (they don't get translated), results may not be very good.

2.6.2. Integration of new concepts, names

Bloggers live and write in a very dynamic world. New concepts, persons, products, places can trend over night. Learning cycle of a machine translation service has to be short enough, so that these new concepts/names get incorporated into translation service as soon as possible, so that they get translated appropriately and possibly improve the results from recommendation engine.

Because Accurat CC-enhanced method depends on news crawling, extracting parallel phrases and training translation workers on these data, the learning cycle is longer than ideal (daily integration), but it is still fast enough.

2.7. Configuration

Translation service has several parameters, which have to be set in a translation request: language pair, translation method and translation token.

3. Implementation

Web authoring applications are software applications that enable users to develop a web site in a desktop publishing format. Software generates required HTML, user just has to enter contents. One type of where user has to deal with the content only, and not so much about underlying technologies.

Most popular blogging platforms are WordPress[1], Blogger[2] and TypePad[3]. They offer .

3.1. Technologies in web authoring tools

The most frequently used technologies in the domain of web authoring are JavaScript, (especially jQuery and Ajax), PHP, and Python/Django, because they are relatively easy to use, they are lightweight and can run on every system. These technologies were also used for implementation of Zemanta widget and for Accurat version of Zemanta WorpdPress plugin. They also provide simple ways of implementing clients for REST services.

3.2. Service requests

For each translation a request has to be created with appropriate parameters including tokenized text, language pair (source and destination language), translation method and translation token for authorization.

3.2.1. Ease of use

Service request should be very easy to construct, so that a client can be implemented with a combination of jQuery, Ajax, and PHP or python. Translation service API offers XML-RPC interface to serverland dashboard[4].

One of potential problems at this point is tokenization, which has to be done on the client side. This is just another layer of functionality that has to be taken care of.

3.2.2. Price

Several web services, e.g. Zemanta or OpenCalais among others, offer their service for free under certain terms and with limitations. Use of Zemanta API[5] is limited to 1000 calls per day, while OpenCalais[6] is limited to 50000 calls per day. On the other hand, Google Translate API[7] charges for translations and language detection based on usage (in millions of characters).

3.3. Zemanta widget with integrated Accurat MT

We implemented a plugin for WordPress with integrated Accurat MT service. Typical workflow of blogging using Zemanta (without translation service) is shown on Figure 1 (a). User types text in online editor on preferred blogging platform and in the meanwhile Zemanta widget checks how much text is written. When 300 characters are reached, text is sent to Zemanta recommendation engine, which returns related content and displays it on a widget. After that Zemanta is periodically checking whether new characters have been writted and updates related content (related images, related articles) on the widget.

[1] Wordpress blogging platform: http://wordpress.com/

[2] Blogger blogging platform: http://www.blogger.com

[3] TypePad blogging platform: http://www.typepad.com/

[4] Mt serverland code on github – XML-RPC interface: https://github.com/cfedermann/mt-serverland/blob/master/serverland/dashboard/api/xmlrpcserver.py

[5] Zemanta developer's documentation page: http://developer.zemanta.com/docs/

[6] OpenCalais developer's page: http://www.opencalais.com/about/developer

[7] Google Translation API pricing page: https://developers.google.com/translate/v2/pricing

While Zemanta works best with English texts, it also returns content for texts in other languages, but in most cases not as good as if the text is in English (Figure 1, b). Zemanta always recommends related articles, but some of them are actually related to the text and others are considered noise.

Zemanta widget with integrated translation service added another step in this workflow (Figure 1, c). Again the length of the text gets checked and when it riches 300 characters, the text is sent to Accurat MT and Zemanta widget periodically checks, whether translated text is ready. When it is, it sends text translated in English to the Zemanta recommendation engine, which then returns related articles. In this case we usually get better, more relevant related articles and related images.

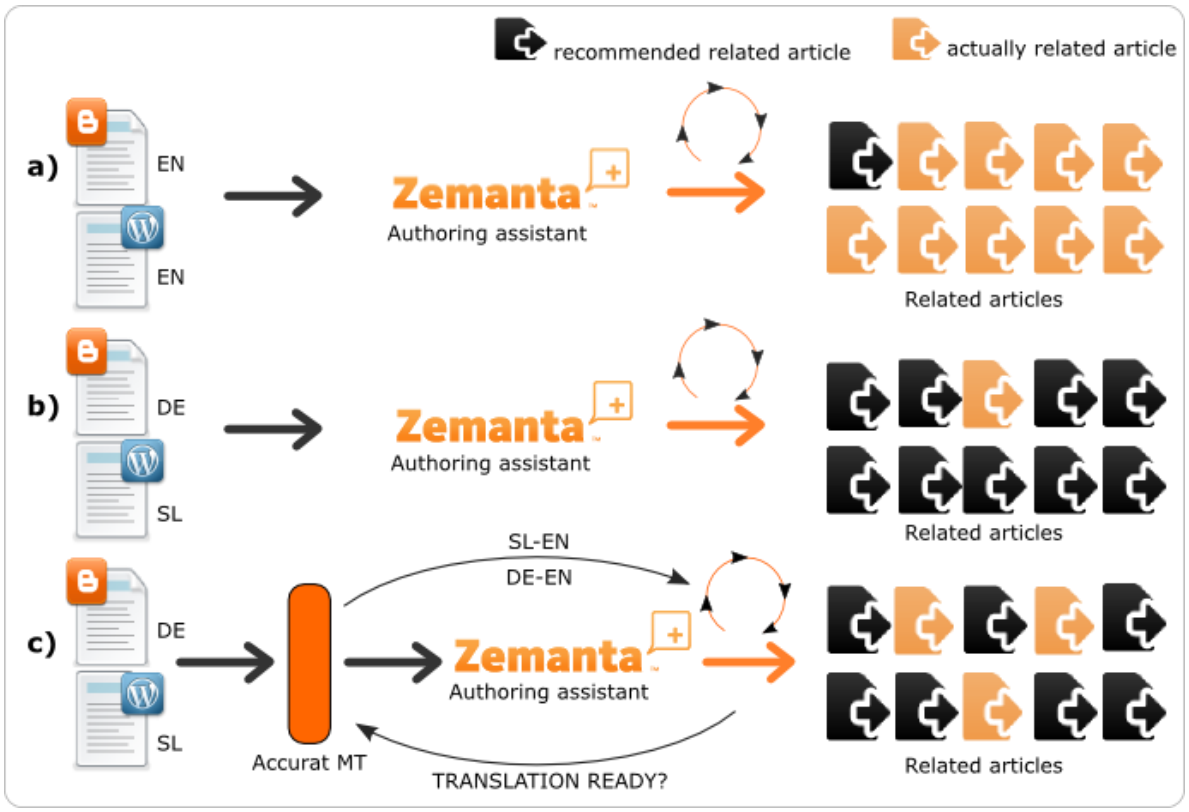

Figure 1: Workflow using Zemanta and Accurat MT

3.3.1. Tokenization on client side

Tokenization of text is done by Zemanta widget on our servers, because it was too complicated to implement this in on user's (client) side.

3.3.2. Authorization - requesting token

We were provided with translation token for demonstration purposes and therefore this plugin is not publicly available.

4. Evaluation

We had following objectives in our evaluation task. First we wanted to evaluate the possibility of using Accurat MT for web authoring. We wanted to find out if recommendation system with integrated machine translation can help authors be more productive and effectivein their writing. In our case, where translation is used as an intermediate step, this means more really related articles to include in the blog post. Our other objective was to

implement Accurat MT methods in our web-authoring application and use Accurat MT for translations of blog posts in Slovenian, Gernam or Croatian language into English.

Our main goal was to obtain relevant data on the plausible uses of machinte translations in Zemanta's recommendation engine by using machine translation as an intermediate technology.

The quality of machine translation can be viewed from different aspects. e.g. from human translator's view, who has to deal with translated text directly, or from Zemanta's view, where translation is only an intermediate step.

In our case user is not interested in translated text *per se*, but in results obtained from Zemanta's engine in form of related articles. We assumed that using machine translation service will make a difference and our null hypothesis was that results returned for original texts and translated texts will not differ significantly.

Evaluation was carried out in three parts:

- Evaluation with Zemanta's internal tool Dash
- Detailed analysis of randomly selected files
- Use case using blogging platform

4.1. Evaluation with internal tool Dash

In this evaluation part we used sets of 100 texts for each language pair (Table 1). Internal tool Dash displays related articles and provides simple graphical user interface for human evaluators.

Table 1: Evaluation sets of texts

Language pair	Number of files	Avg. text length (words)
SL-EN	100	**238,8**
DE-EN	100	**242,7**
HR-EN	100	**202,7**

Evaluation scenario we used:

- Translations (SL-EN, DE-EN, HR-EN) were obtained from Accurat MT
- Translatated texts were used as **input to Zemanta recommendation engine**
- Engine returned 10 related articles for each translated text
- Recommendations were evaluated by human evaluators
- Evaluation results were given as precision@10 metric

The evaluation process was twofold: first we evaluated texts in original language and then we repeated the process with texts translated into English using baseline machine translation. In the first part original texts have been fed to Zemanta's recommendation engine, which provided 10 related articles per text. Each of the articles was manually checked by human evaluators, who decided whether suggested article is actually related to the content (text analyzed) in question or not by assigning it a score between 0 (a blogger would definitely not use it) and 3 (a blogger would definitely use it). After evaluators assigned scores to all related

articles for all of the texts, we calculated precision to estimate the quality of machine translation methods. Figure 2 shows translated text (on left) and recommended related articles (with scores) on the right.

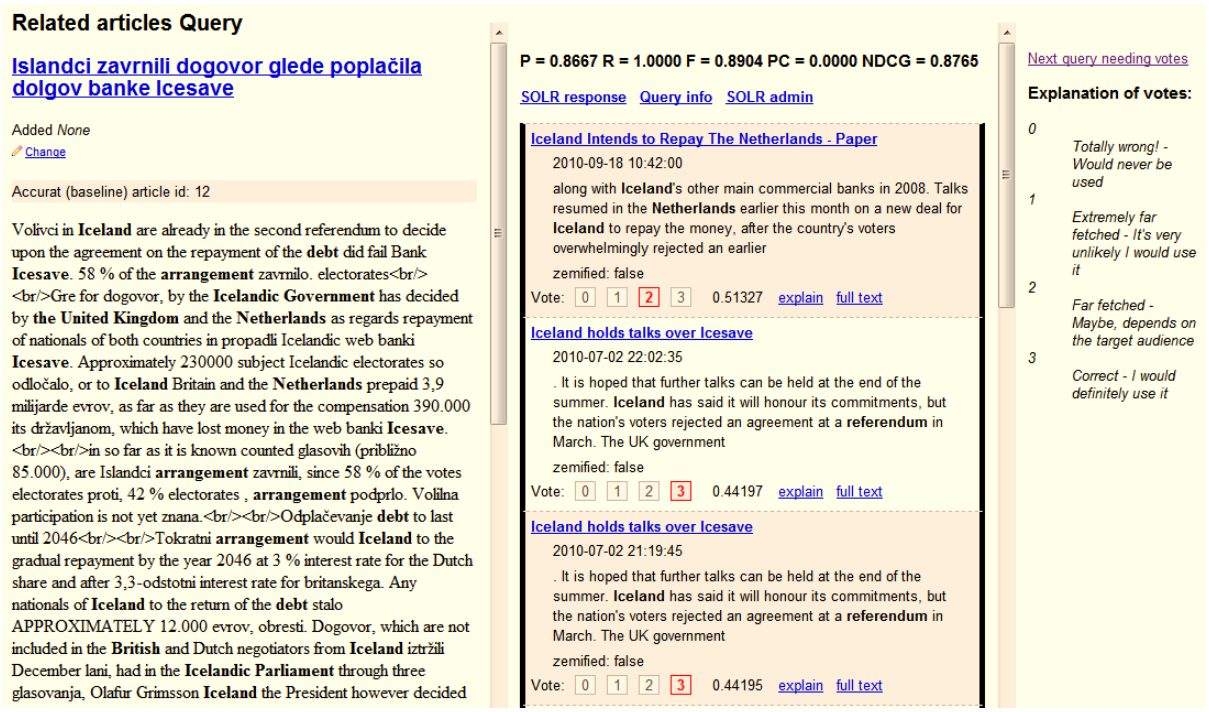

Figure 2: Evaluation in internal tool Dash

When CC-enhanced method was available, we repeated this process to obtain evaluation results and to compare them with results from baseline MT (Figure 3).

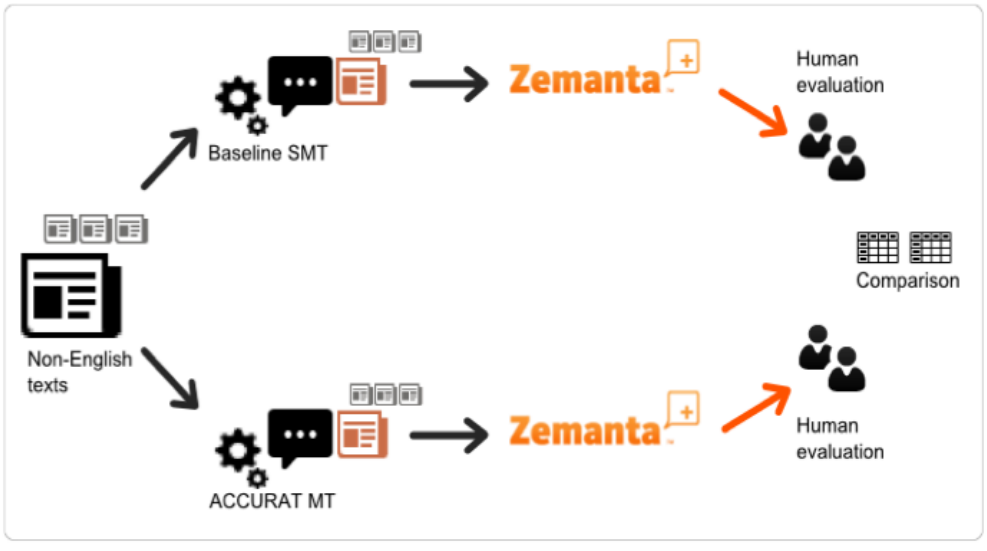

Figure 3: Evaluation process with internal tool Dash

4.1.1. Evaluation metrics

For evaluation we used evaluation metrics Precision@10, which considers only top 10 relevant documents with highest precision score. Figure 4 shows user interface in Dash with precision scores for each evaluated text.

$$precision = \frac{\{relevant\_documents\} \cap \{retrieved\_documents\}}{\{retrieved\_documents\}}$$

Article title	P	R	F	Results voted	Results total	PC	NDCG	ms
Lady Gaga bo kot urednica opozarjala na pravice istospolnih	0.6667	1.0000	0.7143	10	10	0.0000	0.6667	525
Nato se opravičuje s stisnjenimi zobmi	0.0333	1.0000	0.0413	10	10	0.1000	0.0273	1211
Vettel spet v najboljšem položaju	0.6000	1.0000	0.6522	10	10	0.1000	0.5398	399
Četrti zaporedni poraz Lakersov	0.2667	1.0000	0.3125	10	10	0.0000	0.2438	1250
ZDA zaskrbljene zaradi nasilja nad ženskami in geji v Sloveniji	0.6000	1.0000	0.6522	10	10	0.0000	0.6215	1634
Minister: Izrael ne bo dopuščal raket iz Gaze	0.7000	1.0000	0.7447	10	10	0.0000	0.6940	910
Bodo poškodovano Fukušimo razgradili v desetih letih?	0.3333	1.0000	0.3846	10	10	0.0000	0.3333	1034
Nemški gospodarstveniki iščejo slovenska podjetja	0.0000	1.0000	0.0000	10	10	0.0000	0.0000	1194
Že skoraj 10.000 kazni zaradi vožnje brez vinjete	0.0000	1.0000	0.0000	10	10	0.0000	0.0000	1078
Jaz sem četrti	0.8333	1.0000	0.8621	10	10	0.0000	0.8411	838
Islandci zavrnili dogovor glede poplačila dolgov banke Icesave	0.8667	1.0000	0.8904	10	10	0.0000	0.8765	988

Figure 4: Internal tool dash with precision for each article

4.1.2. Intermediate results

Evaluation was performed several times during the project. First results for Slovenian news texts in September 2011 were quite promising. In January 2012 we evaluated two new sets of texts, this time blog posts, for Slovenian and German. Results were even better (Figure 5).

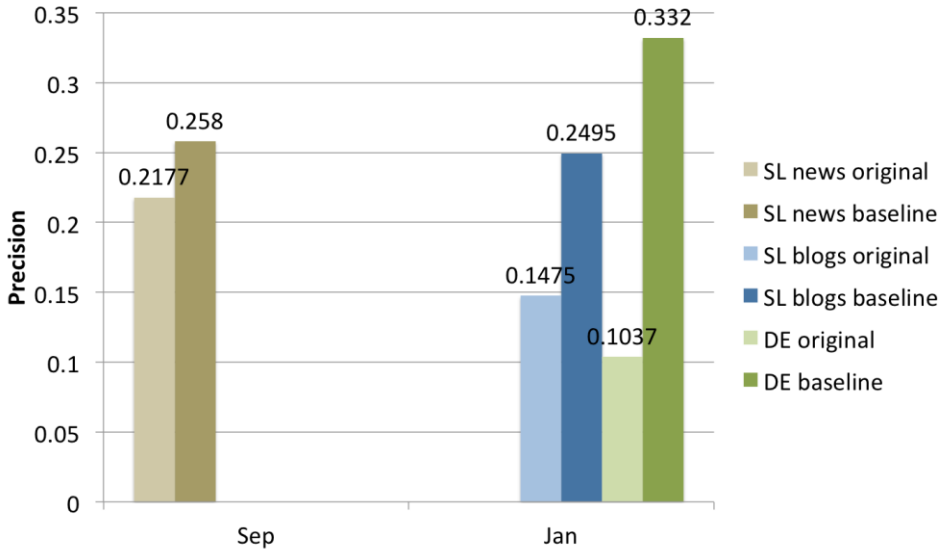

Figure 5: Intermediate results for SL-EN, DE-en language pars using baseline method

4.1.1. Final evaluation

Final evaluation was performed from April till June 2012. In all cases improvement in precision was calculated from average precision.

Results for Slovenian-English pair

If we take a look at the results for Slovenian-English language pair (Table 1), we can see, that using baseline MT in comparison to original texts improved precision for 11% and using CC-enhanced MT improved it for 15%.

Table 2: Evaluation results (precision) for SL-EN pair on three different sets of texts

Dataset	Evaluator 1	Evaluator 2	Average
Original	0,159	0,148	0,153
Baseline	0,281	0,249	0,265
CC-enhanced	0,323	0,277	0,299

Results for German-English pair

Results for using Accurat MT for German texts shows even greater improvement: 20% for baseline MT and 24% for CC-enhanced MT in comparison to original texts (Table 3).

Table 3: Evaluation results (precision) for DE-EN pair on three different sets of texts

Dataset	Evaluator 1	Evaluator 2	Average
Original	0,179	0,104	0,141
Baseline	0,355	0,332	0,344
CC-enhanced	0,405	0,354	0,379

Results for Croatian-English pair

For Croatian-English language pair we were able to evaluate baseline MT only, which improved results for 11% (Table 4).

Table 4: Evaluation results (precision) for HR-EN pair on three different sets of texts

Dataset	Evaluator 1	Evaluator 2	Average
Original	0,212	0,191	0,201
Baseline	0,313	0,314	0,314
CC-enhanced	x	x	x

4.1.2. Sumary

Evaluation results for all three language pairs and both translation methods are shown on Figure 6.

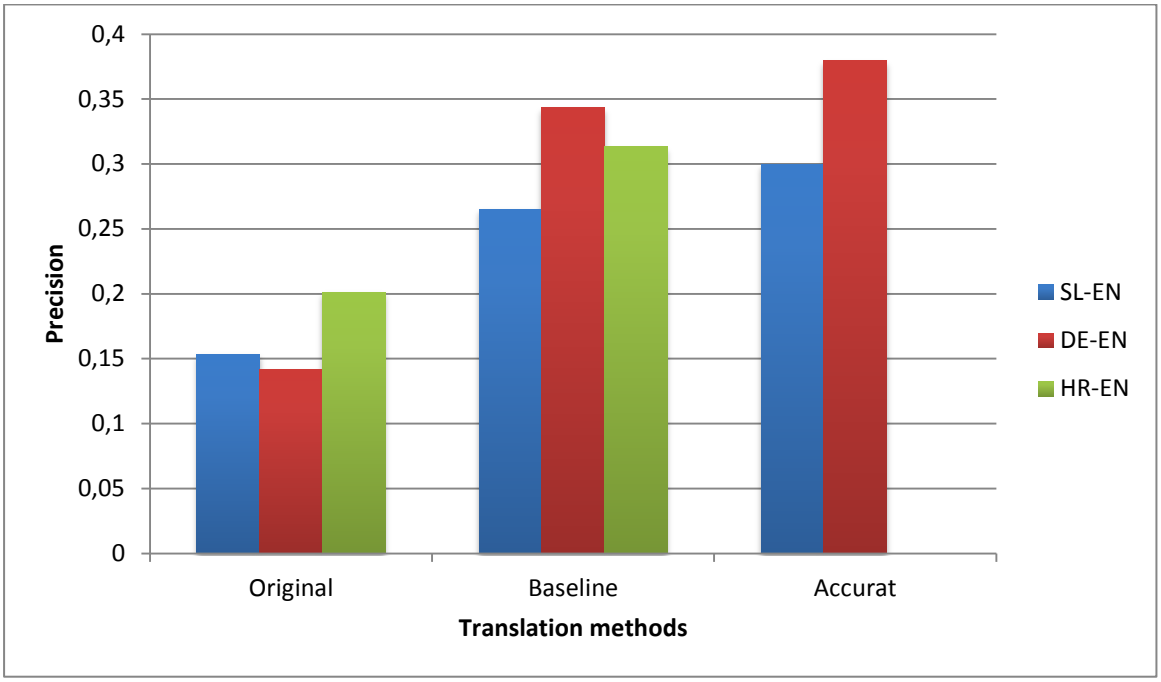

Figure 6: Summary of evaluation results (average precision) in internal tool Dash

Our null hypotheses stated that *evaluation results for original texts do not differ significantly from results for baseline method and CC-enhanced method*. We tested hypotheses using unpaired t-test. Values used in tests and P-values obtained are summarized in Table 5 and Table 6.

Table 5: Mean, standard deviation for original texts

	SL	DE	HR
Mean	0,153366	0,141333	0,201166
STD	0,210006	0,181537	0,306484

For both methods and all langages difference between results for original texts and translated texts were significant on 95% confidence interval.

Table 6: Mean, standard deviation and P-value for baseline method and cc-enhanced method in comparison to original texts

	Baseline			CC-enhanced	
	SL	DE	HR	SL	DE
Mean	0,265151	0,343501	0,313498	0,30123	0,381206
STD	0,243702	0,247905	0,312261	0,242878	0,292788
P-value	0,0006	0,0001	0,0110	0,0001	0,0001

4.2. Detailed analysis

In this part of evaluation we randomly selected 10 texts from each language and analyzed them for translation quality in terms of % of translated/not translated words. We also measured the time needed for translation, but this was done on the whole set of 100 texts.

4.2.1. Translation quality

Recommendation engines are (at least to some point) dependent on keywords. If keywords are not good enough, recommendations from engines may not be good either. Because we didn't know, which keywords were selected by Zemanta's recommendation engine, we just checked the percentage of translated words in the texts for baseline method (Table 7) and CC-enhanced method (Table 8). Tables include average number of words in original files, average number of words in translated files, percentage of translated words and % of not translated words. Numbers were excluded from % of not translated words.

Baseline method does not perform well on Slovenian texts, because only 59% of words have been translated, but according to evaluation results this was still good enough to more relevant recommended related articles.

Table 7: Translation quality for baseline method

Language pair	Avg. words - original	Avg. words – translation	% translated words	% not translated
SL-EN	238,8	232	59	40
DE-EN	242,7	209,8	73	26
HR-EN	202,7	183,8	73	24

Table 8: Translation quality for CC-enhanced method

Language pair	Avg. words - original	Avg. words – translation	% translated words	% not translated
SL-EN	238,8	225,2	76	23
DE-EN	242,7	217,1	74	24
HR-EN	X	X	X	X

4.2.2. Translation time

We also measured time needed for translations of all 100 texts for each language pair. Timings are collected in Table 9 for baseline translation method and in

Table 10 for CC-enhanced method.

Table 9: Translation time for baseline method

Language pair	Avg translation time (sec)	Min time (sec)	Max time(sec)
SL-EN	111,98	61,56	365,05
DE-EN	172,71	92,16	273,62

Language pair	Avg translation time (sec)	Min time (sec)	Max time(sec)
HR-EN	78,92	31,70	122,47

Table 10: Translation times for CC-enhanced MT

Language pair	Avg translation time	Min	Max
SL-EN	133,99	30,95	423,88
DE-EN	186,98	122,06	304,94
HR-EN	x	x	x

4.3. Use case: using Zemanta widget with integrated translation service in blogging platform

To demonstrate Accurat machine translation in the real blogger's environment we installed WordpPress plugin for Zemanta widget with integrated translation on one blog and plugin for Zemanta widget withoud translation service on another one.

We entered text from news in Slovenian (describing latest scandal regarding weapons in USA) into both editors and used Zemanta's widget to get related articles. Recommended related articles for blog using Zemanta widget with integrated translation are shown on Figure 7 and related articles from "normal" Zemanta widget are shown on Figure 8. Although recommended images are almost the same, they show a person (Eric Holder). Because this person's name doesn't get translated and because it represents a named entity, it is present in both widgets.

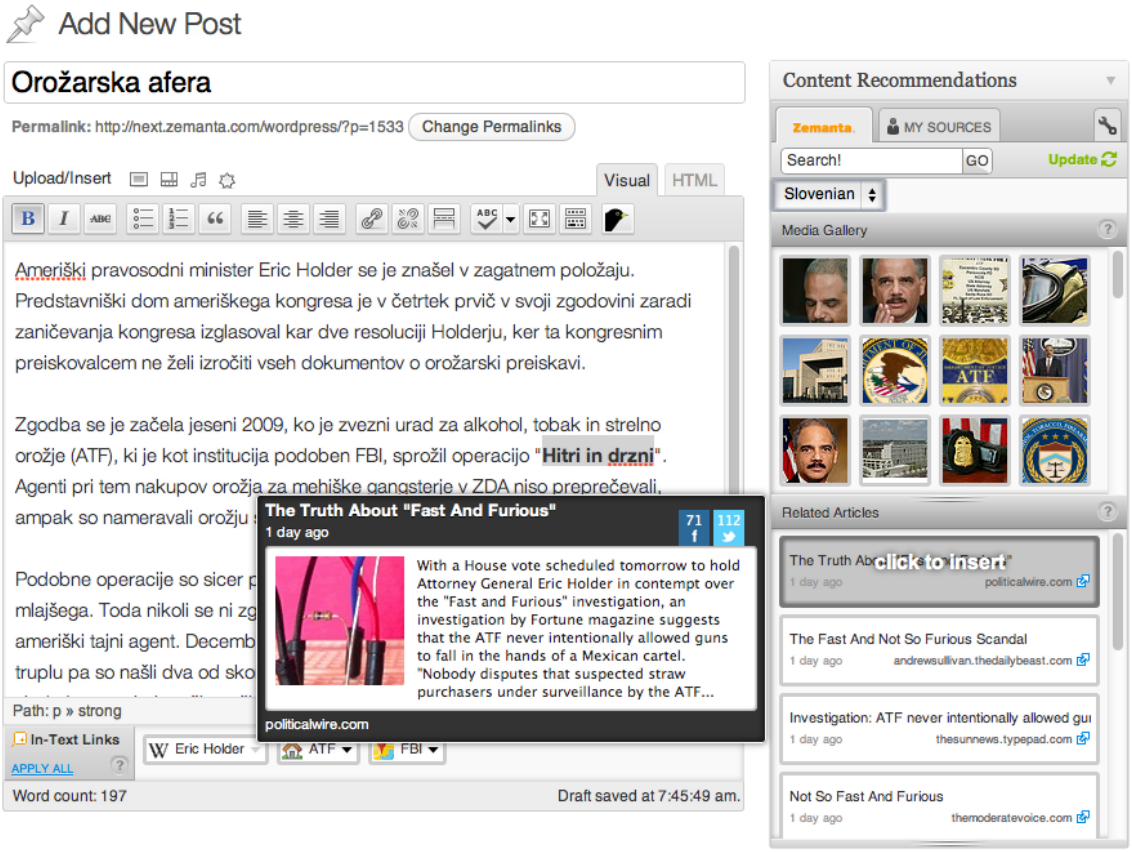

Figure 7: Recommended related articles for German text using Zemanta widget with translation

Figure 8: Recommended related articles for German text using Zemanta widget without translation

5. Conclusion

Evaluation results have shown than machine translation represents new opportunities for web authoring as an intermediate step between texts in languages other than English and recommender service optimized for English language.

Both ACCURAT translation methods – baseline and CC-enhanced method - fulfill all the basic requirements we defined in Section 2 - General Requirements. They also significantly improved the relatedness of recommended articles.

6. References

1. **incite, NM.** Buzz int he Blogosphere: Millions more bloggers and blog readers. *NM incite, know the customer.* [Online] http://nmincite.com/buzz-in-the-blogosphere-millions-more-bloggers-and-blog-readers/.

2. **Technorati.** *Technorati.* [Online] 2012. [Cited: 06 20, 2012.] http://technorati.com/social-media/article/state-of-the-blogosphere-2011-introduction/.

7. List of tables

8. List of figures