Investigação do uso de *word embeddings* para cálculo de similaridade em memórias de tradução

1st Karina Mayumi Johansson
Departamento de Computação – DC
Universidade Federal de São Carlos – UFSCar
São Carlos, Brasil
kahjohansson@gmail.com

2nd Helena de Medeiros Caseli
Departamento de Computação – DC
Universidade Federal de São Carlos – UFSCar
São Carlos, Brasil
helenacaseli@ufscar.br

***Palavras-Chave—Word Embeddings*, Memórias de Tradução, Ferramentas CAT, Português do Brasil, Inglês**

I. Introdução

Os avanços tecnológicos que ocorreram nos últimos anos relacionados à popularização da internet tornaram possível a disponibilização cada vez maior de conteúdo multilíngue, principalmente em sites e repositórios de informação acessíveis e editáveis por todos, como as wikis. Nesse cenário, a tradução desse conteúdo (para entendimento ou disponibilização de uma versão em outro idioma) tornou-se uma demanda não apenas de pesquisadores, cientistas e membros do governo, mas do público em geral. A internet proporcionou às pessoas o acesso à informação de um modo nunca antes visto pela sociedade e a disponibilidade crescente de informação multilíngue se contrapõe à dificuldade dos usuários da internet, espalhados por todo o mundo, em entender esse conteúdo.

Embora existam, atualmente, diversos sistemas e ferramentas para a tradução automática (TA), comerciais ou disponíveis online, os mesmos ainda não são capazes de produzir TA de qualidade comparável a do humano, em domínios irrestritos.

Além dos sistemas completamente automáticos, devido à qualidade contestável das traduções por eles produzidas, outras frentes de pesquisa se desenvolveram com o intuito de criar ferramentas e recursos para auxiliar o humano na tarefa de traduzir ou de editar a tradução antes, durante ou depois de sua realização.

Especificamente em relação à primeira frente, é nela que se inserem as ferramentas investigadas neste trabalho, as *Computer-Assisted Translation* (CAT) *tools*, que utilizam como principal recurso as Memórias de Tradução (MT). Uma MT pode ser entendida como um repositório que armazena os segmentos originais (geralmente frases) e as respectivas traduções humanas, com o objetivo de serem manipulados e reaproveitados em traduções futuras. Com o passar do tempo de uso, a MT agrupa um grande conjunto de segmentos acompanhados de suas respectivas traduções e esses podem ser recuperados pela ferramenta CAT por meio de um casamento (*matching*), completo ou parcial, sempre que um segmento igual ou semelhante aparecer novamente. Assim, uma MT pode ser considerada um córpus paralelo, pois contém segmentos alinhados de textos já traduzidos, que têm a função de serem reaproveitados em uma futura tradução.

Entre as principais vantagens do uso de memórias de tradução estão: a consistência e a agilidade. As MT asseguram que os documentos traduzidos são consistentes, incluindo terminologia, estruturas, expressões e definições comuns. Além disso, elas também aceleram o processo de tradução já que dividem com o tradutor humano a pesada carga de tradução recordando para ele o que já foi previamente traduzido e deixando-o focado na tradução de novos trechos.

Ferramentas CAT podem oferecer, ainda, funcionalidades extra, como: análises estatísticas, exportação e importação de MT, conversão de formatos, trabalho colaborativo e alinhamento. Por meio do alinhamento é possível estabelecer a correspondência entre segmentos (ou palavras) de um arquivo original com os segmentos (ou palavras) de um arquivo que foi traduzido sem o uso das MT permitindo, assim, a criação "automática" de uma MT ou um glossário. Apesar de serem bastante úteis, essas funcionalidades extras não são o foco principal deste trabalho.

Neste trabalho apresentamos um projeto cujo objetivo é investigar a aplicabilidade de *word embeddings* para implementar a funcionalidade principal relacionada ao uso de MT em uma ferramenta CAT, que é o casamento (*matching*) entre segmentos da sentença sendo traduzida e os segmentos presentes na MT. A estratégia tradicionalmente usada para implementar o casamento é considerar a intersecção (ou sobreposição) nas sequências de palavras (n-gramas) presentes nos segmentos de texto em comparação, o que pode ser descrito como casamento de n-gramas. Essa estratégia tradicional, contudo, não é capaz de capturar similaridade semântica além do nível trivial [2].

Como alternativa ao casamento de n-gramas, o projeto apresentado neste trabalho propõe investigar como as *word embeddings* podem ser usadas para encontrar a similaridade entre segmentos de uma MT. As *word embeddings* (ou vetores de palavras) são representações vetoriais de palavras, geradas de modo não supervisionado a partir de córpus. Elas representam as palavras em um espaço vetorial no qual a similaridade semântica entre duas palavras é calculada com base na proximidade dos vetores que as representam [2].

Neste trabalho, portanto, apresenta-se um projeto de investigação do uso de *word embeddings* mono e bilíngues

para implementar a funcionalidade de casamento de segmentos de texto fundamental nas ferramentas CAT. Para tanto, a seguir, são descritos brevemente os dois tópicos principais desta pesquisa: memórias de tradução e *word embeddings*. Em seguida, são descritos os objetivos e a metodologia a ser seguida, bem como as duas ferramentas CAT selecionadas como candidatas a alteração nesse projeto. Por fim, são apresentados os resultados esperados e as possíveis contribuições do projeto em questão.

II. Memórias de Tradução

Segundo [5], as memórias de tradução (MT) são uma das principais fontes de conhecimento que dão suporte à tradução humana nas ferramentas CAT. Uma MT é uma base de dados que armazena segmentos fonte e alvo chamados de unidades de tradução (do inglês, *translation units* ou TU). Esses segmentos podem ser fragmentos sub-sentenciais, sentenças completas ou mesmo parágrafos em duas línguas e, idealmente, eles são traduções perfeitas uns dos outros.

Para que esses segmentos sejam usados em uma ferramenta CAT é necessário calcular uma pontuação de casamento entre uma sentença de entrada sendo traduzida e o lado fonte de cada TU armazenada na MT. Se a pontuação atingir um valor mínimo, o lado alvo da TU é sugerido como uma tradução para o usuário [5].

Novas TUs são inseridas na MT a todo momento, ou seja, sempre que o usuário realiza a tradução de um segmento fonte gerando um segmento alvo equivalente para ele, este par (essa TU) pode ser armazenado na MT para uso futuro. Esse crescimento constante da MT garante sua utilidade para o tradutor humano que não precisará, no futuro, re-traduzir segmentos previamente traduzidos garantindo a consistência nas traduções e agilizando seu trabalho. Como bem colocado por [5], juntamente com a quantidade, a qualidade das TUs armazenadas é um fator determinante na utilidade da MT.

Exemplos de sistemas de MT são OmegaT[1], Wordfast[2], Déjà Vu da Atril Solutions[3], memoQ[4] e SDL Trados[5] entre outros.

A Figura 1 ilustra o uso de uma MT em uma ferramenta CAT, o SDL Trados. Na tabela presente na parte inferior da figura são apresentados, na coluna da esquerda, os segmentos do texto fonte e, na coluna da direita, as traduções correspondentes a cada segmento. Na parte superior da figura são apresentadas as semelhanças e diferenças do segmento atual do texto fonte com o segmento da MT com melhor pontuação de casamento.

Ainda na parte superior, à direita, é apresentada a tradução recuperada da MT e sua pontuação de casamento. A primeira sentença teve uma pontuação de casamento (do inglês, *context match* ou CM) de 100% uma vez que seu contexto era semelhante ao da TU. Nesse caso, o usuário optou por recuperar a tradução da TU e utilizá-la sem modificações. Para as segunda e terceira sentenças, é possível observar que os valores de CM foram, respectivamente, 88% e 77%. Nesses casos, o usuário optou por realizar modificações na tradução recuperada.

Assim, a partir das informações apresentadas na ferramenta, o tradutor humano verifica a sentença fonte selecionada e caso tenha sido encontrada uma TU semelhante na MT, ele pode optar por: (1) utilizar a tradução da TU e modificá-la se necessário, ou (2) ignorar a tradução sugerida e criar uma nova tradução para o segmento. Ao confirmar a tradução e prosseguir para o novo segmento, a nova TU é guardada na MT, e pode ser utilizada nos demais segmentos do texto, assim como em outros projetos, por meio da importação dessa MT. Deste modo, o tradutor segue esse processo iterativo até o fim do documento.

III. Word Embeddings

As *word embeddings* (ou vetores de palavras) são representações vetoriais de palavras, geradas de modo não supervisionado a partir de córpus, que representam as palavras em um espaço vetorial. A similaridade semântica entre duas palavras é, então, calculada com base na proximidade dos vetores que as representam [2].

Essas similaridades e disparidades semânticas também podem ser estendidas para duas ou mais línguas. Em [3], por exemplo, os autores propõem a construção de modelos de língua monolíngues para inglês, espanhol e tcheco usando grandes córpus e, em seguida, usam um dicionário bilíngue pequeno para aprender uma projeção linear entre os vetores que representam cada língua. Desse modo, a tradução de uma palavra fonte é obtida projetando seu vetor para a língua alvo e buscando pelo vetor mais similar na língua alvo.

Exemplos de ferramentas que geram *word embeddings* são word2vec[6], GloVe[7], fastText[8] e MUSE[9].

Como as *word embeddings* capturam as similaridades sintáticas e semânticas em uma ou várias línguas, elas têm sido aplicadas, entre outros, para encontrar a similaridade entre textos. Em [1], por exemplo, foram usadas *word embeddings* bilíngues para recuperar sentenças semanticamente similares. Foram utilizados três pares de línguas: inglês-espanhol, inglês-italiano e inglês-croata. Como explicado pelos autores, a similaridade semântica em textos é estimada com base no quanto dois textos estão semanticamente relacionados ou associados, o que varia desde a equivalência semântica (na qual o significado dos dois textos é exatamente o mesmo) até a completa disparidade (na qual o significado de um texto está completamente dissociado do significado do outro). Assim, segundo esses autores, a similaridade semântica entre dois textos é dada pela sobreposição de significado entre os dois

[1]Disponível em: https://omegat.org/. Acesso em: 28 maio 2019.
[2]Disponível em: https://www.wordfast.net/. Acesso em: 28 maio 2019.
[3]Disponível em: https://atril.com/. Acesso em: 28 maio 2019.
[4]Disponível em: https://www.memoq.com/en/. Acesso em: 28 maio 2019.
[5]Disponível em: https://www.sdltrados.com/. Acesso em: 28 maio 2019.
[6]Disponível em: http://deeplearning4j.org/word2vec. Acesso em: 29 maio 2019.
[7]Disponível em: https://nlp.stanford.edu/projects/glove/. Acesso em: 29 maio 2019.
[8]Disponível em: https://fasttext.cc/. Acesso em: 29 maio 2019.
[9]Disponível em: https://github.com/facebookresearch/MUSE. Acesso em: 29 maio 2019.

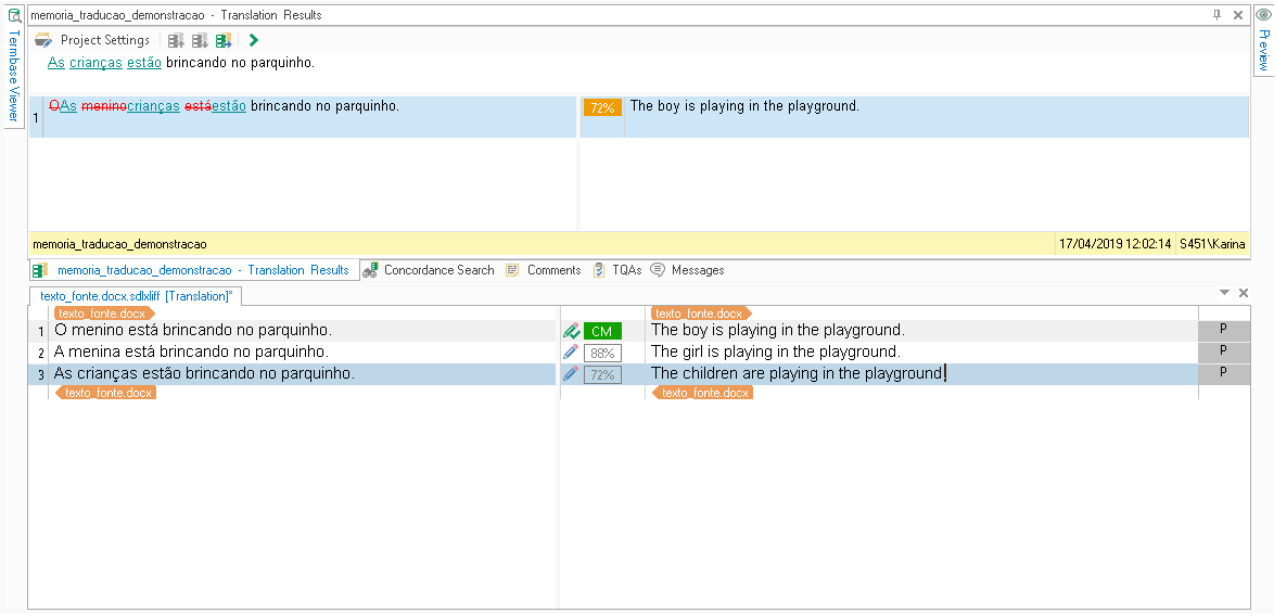

Figura 1. Exemplo de utilização de sistema de memória de tradução.

textos, por exemplo, as sentenças "o gato está ronronando" e "o cachorro está latindo" são mais relacionadas do que "a tartaruga está correndo".

No âmbito das MTs, *word embeddings* bilíngues do par de línguas inglês-italiano foram usadas por [5] como uma das *features* no processo de limpeza das MTs, no qual TUs de baixa qualidade foram eliminadas da MT.

IV. Objetivos e Metodologia

A partir da contextualização apresentada anteriormente, pode-se estabelecer como objetivo do projeto aqui apresentado: verificar a aplicabilidade de *word embeddings* no cálculo da similaridade entre segmentos de uma sentença sendo traduzida e os segmentos de uma MT.

Assim, as seguintes questões de pesquisa são levantadas para este trabalho:

1) O uso de *word embeddings* para casamento de segmentos de uma sentença sendo traduzida e os segmentos de uma MT (TUs) é melhor do que a estratégia tradicional baseada em n-gramas?
2) O uso de *word embeddings* bilíngues em conjunto com *word embeddings* monolíngues é efetivo?

Para atingir o objetivo apresentado anteriormente e responder às questões de pesquisa levantadas, as seguintes etapas serão realizadas:

1) Levantamento das principais ferramentas CAT de código aberto para seleção daquela que será alterada nesse projeto,
2) Levantamento das memórias de tradução para o português do Brasil e o inglês que poderão ser usadas como córpus base para a avaliação,
3) Levantamento das *word embeddings* disponíveis para o português do Brasil e o inglês, bem como o estudo das principais ferramentas de geração de *word embeddings* mono e bilíngues, caso seja necessário gerar *word embeddings* específicas para esse projeto,
4) Proposição e implementação da estratégia de cálculo de similaridade de segmentos de MT usando separada e conjuntamente *word embeddings* monolíngues e bilíngues,
5) Avaliação das estratégias propostas com base na comparação da qualidade do *matching* no sistema *baseline* (versão do sistema CAT sem qualquer alteração) e nas versões do sistema nas quais será implementado o *matching* usando *word embeddings* mono e bilíngues separada e conjuntamente.

O projeto sobre o qual trata este trabalho está no início de seu desenvolvimento e, até o presente momento, apenas a primeira atividade foi finalizada, para a qual os resultados são apresentados na seção V.

V. Ferramentas CAT

A primeira atividade do projeto ao qual este trabalho se refere trata do levantamento e estudo das ferramentas CAT de código aberto disponíveis livremente. Nesse sentido, como resultado de um primeiro levantamento foram selecionadas as seguintes ferramentas: MateCat [7] e OmegaT [8].

As ferramentas citadas possuem recursos em comum, como: (i) utilização de memórias de tradução com *fuzzy matching*[10] e (ii) pré-tradução por meio de *plugins* que acessam tradutores automáticos como Apertium[11] e Google Tradutor[12].

[10] *Fuzzy matching* é um casamento com taxa de similaridade menor que 100% entre o segmento a ser traduzido e um segmento da memória de tradução. Cada ferramenta CAT possui um valor padrão de limitante inferior, que pode ser alterado pelo usuário.

[11] Disponível em: https://www.apertium.org/. Acesso em: 29 maio 2019.

[12] Disponível em: https://translate.google.com/. Acesso em: 29 maio 2019.

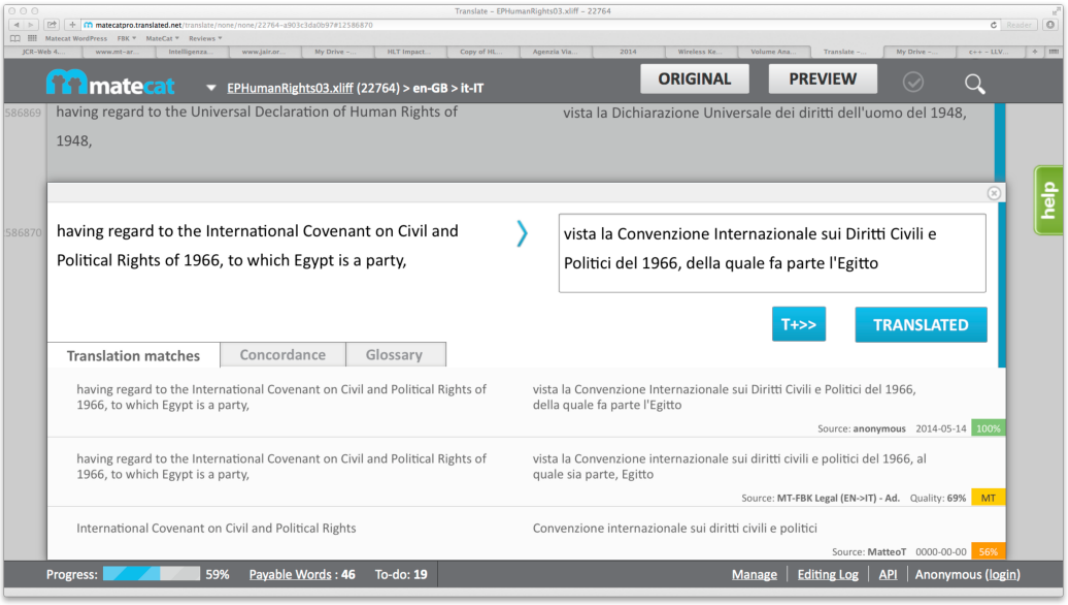

Figura 2. *Interface* da ferramenta CAT MateCat. [6]

Uma diferença existente é que OmegaT foi desenvolvido em plataforma *desktop* e MateCat em plataforma *web*.

MateCat é um acrônimo para *Machine Translation Enhanced Computer Assisted Translation*. Trata-se de uma ferramenta CAT com funcionalidades de: (i) pós-revisão, (ii) *outsourcing* de gerenciamento de projetos, traduções e revisões e (iii) trabalho em equipe, por meio de divisão e gerenciamento de projetos. A Figura 2 apresenta sua *interface*, contendo em seu lado esquerdo os segmentos fonte, e à direita, suas respectivas traduções.

OmegaT é uma ferramenta desenvolvida em Java, com *plugins* em diversas linguagens, como: Javascript, Groovy e Python. Seus principais diferenciais são: (i) propagação de correspondência (*match propagation*), ou seja, assim que um segmento é traduzido, a mesma tradução é inserida automaticamente em todos os segmentos idênticos e (ii) suporte de dicionários mono e multilíngue. A Figura 3 apresenta sua tela de tradução, que possui no canto inferior esquerdo os segmentos originais e abaixo de cada um deles, suas traduções. No canto superior esquerdo são apresentadas as *fuzzy matches*, e ao seu lado direito, são apresentadas as sugestões do tradutor automático.

A Tabela I apresenta um comparativo dessas ferramentas. Além da análise de características pontuais, como as listadas na Tabela I, uma análise qualitativa das funcionalidades e facilidade de uso e alteração de código ainda será realizada para determinar qual delas será a escolhida para alteração nesse projeto.

Tabela I
COMPARATIVO DAS FERRAMENTAS CAT

Categorias	**Ferramentas**	
	MateCat	**OmegaT**
Plataforma	*Web*	*Desktop*
Linguagem de programação utilizada	PHP	Java
Licença	LGPL license	GNU General Public License v3.0
MT públicas integradas	MyMemory	MyMemory
Formatos de MT suportados	TMX	TMX, TTX, TXML, XLIFF e SDLXLIFF

VI. CONTRIBUIÇÕES E RESULTADOS ESPERADOS

Este trabalho apresentou um projeto, em estágio inicial de desenvolvimento, que visa investigar a aplicabilidade de *word embeddings* no cálculo da similaridade entre segmentos de uma sentença sendo traduzida e os segmentos de uma memória de tradução.

Ao final desse projeto, espera-se obter uma estratégia para cálculo da similaridade semântica que seja uma alternativa à estratégia de casamento tradicional baseada em n-gramas. Para validar e avaliar a estratégia proposta ela será implementada/incorporada em uma ferramenta CAT de código aberto permitindo casamentos mais semanticamente motivados e, como tal, mais abrangentes.

Embora *word embeddings* tenham sido usados para detecção de similaridade textual [1], [2] e para limpeza de MT [5], não se tem notícia de um trabalho que tenha investigado a aplicação de *word embeddings* mono e bilíngues para o casamento de segmentos nas MTs. Assim, este trabalho surge como a primeira iniciativa de investigação neste contexto.

Uma outra contribuição deste projeto refere-se ao principal idioma sob investigação, o português do Brasil, que ainda é carente de pesquisas e desenvolvimento de recursos/ferramentas

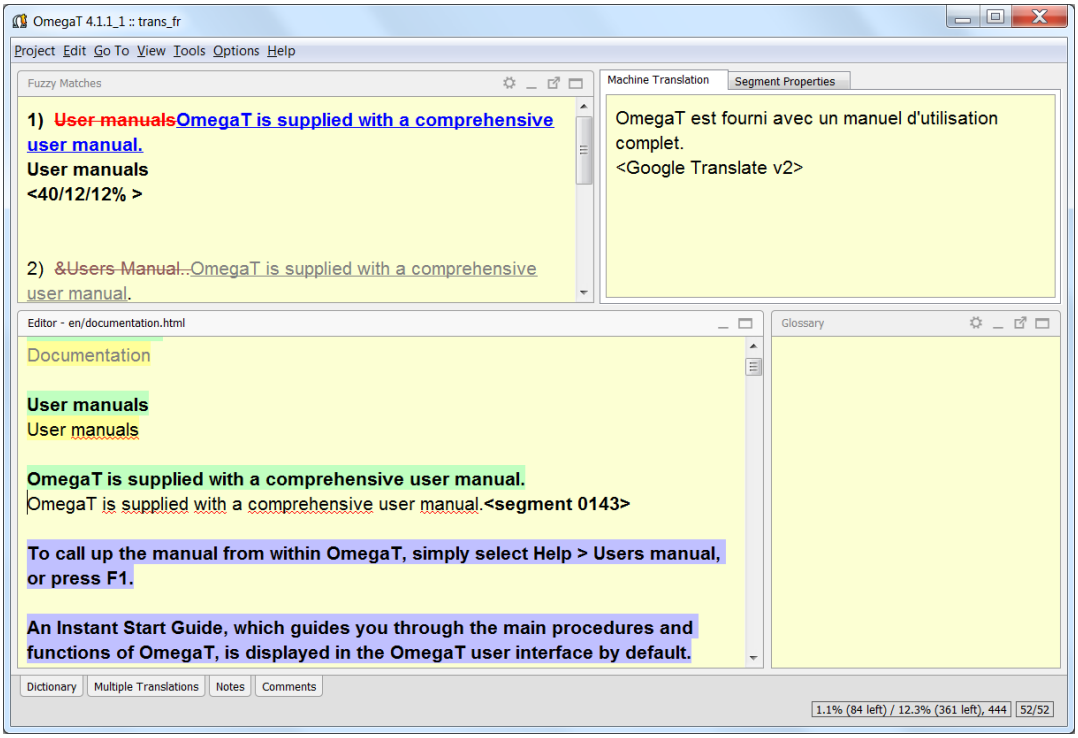
OmegaT 4.1.1_1 :: trans_fr
Project Edit Go To View Tools Options Help
Fuzzy Matches
1) User manualsOmegaT is supplied with a comprehensive user manual.
User manuals
<40/12/12% >
2) &Users Manual..OmegaT is supplied with a comprehensive user manual.
Machine Translation
Segment Properties
OmegaT est fourni avec un manuel d'utilisation complet.
<Google Translate v2>
Editor - en/documentation.html
Documentation
User manuals
User manuals
OmegaT is supplied with a comprehensive user manual.
OmegaT is supplied with a comprehensive user manual.<segment 0143>
To call up the manual from within OmegaT, simply select Help > Users manual, or press F1.
An Instant Start Guide, which guides you through the main procedures and functions of OmegaT, is displayed in the OmegaT user interface by default.
Glossary
Dictionary
Multiple Translations
Notes
Comments
1.1% (84 left) / 12.3% (361 left), 444
52/52

Figura 3. *Interface* da ferramenta CAT OmegaT. [8]

para a tradução (humana e automática) quando comparado ao cenário de outros idiomas, como o inglês.

Referências

[1] G. Glavas, M. Franco-Salvador, S. Ponzetto and P. Rosso, "A Resource-Light Method for Cross-Lingual Semantic Textual Similarity," 2018.

[2] T. Kenter and M. Rijke, "Short Text Similarity with Word Embeddings," in Proc. 24th ACM International on Conference on Information and Knowledge Management (CIKM '15), 2015, pp. 1411-1420.

[3] T. Mikolov, Q. Le and I. Sutskever, "Exploiting Similarities among Languages for Machine Translation," 2013.

[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems (NIPS), 2013, pp. 3111-3119.

[5] M. J. Sabet, M. Negri, M. Turchi and E. Barbu, "An Unsupervised Method for Automatic Translation Memory Cleaning," in Proc. 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 287–292.

[6] M. Frederico et al., "The Matecat Tool," in 25th International Conference on Computational Linguistics, 2014.

[7] "MateCat," matecat.com [Online]. Available: https://www.matecat.com. [Accessed May 29, 2019].

[8] "OmegaT," omegat.org [Online]. Available: https://omegat.org/. [Accessed May 29, 2019].