Traductor estadístico wixarika - español usando descomposición morfológica

Jesús Manuel Mager Hois[1] Carlos Barrón Romero[1] y Ivan Vladimir Meza Ruiz[2]

[1] Universidad Autónoma Metropolitana, Unidad Azcapozalco
[2] Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas

Resumen En este artículo se presenta un traductor automático entre las lenguas español y wixarika, usando traducción estadística y recursos gramaticales complementarios. El wixarika es una lengua indígena hablada en los estados mexicanos de Jalisco, Nayarit, Zacatecas y Durango. Este trabajo se enfoca en dos problemas: la escasa existencia de corpus paralelos y la dificultad de alinear una lengua fusionante (español) con una altamente polisintética (wixarika). En situaciones límites los traductores típicos basados en traducción estadística usan entre 100 y 300 MB de texto alineado. Nuestra propuesta introduce un analizador morfológico que descompone los verbos del wixarika y los expone a la fase de alineamiento.
Palabras Clave: Traducción Estadística Automática, Alineamiento de Lenguas Polisintéticas, Recursos Escasos, Procesamiento de Lenguaje Natural.

1. Introducción

La traducción entre lenguas y la necesidad de comunicación entre las personas se remonta a los origines de nuestra civilización. Hoy con los avances en las Tecnologías de la Información y la Comunicación (TIC) es un tema relevante de investigación interdisciplinaria entre las Ciencias Sociales y la Ciencia de la Computación. El campo semántico lleva, según Tarski, a la indefinibilidad[24] y, por lo tanto, los lenguajes naturales no pueden ser resueltos como lenguajes formales (de la lógica de la ciencia de la computación, por ejemplo, un lenguaje de programación. Pero este no es adecuado para otro fin que no sea el de desarrollar programas de computadora, mientras que las lenguajes naturales, sirven para muchos fines abiertos, comunicación, creación de conceptos y conocimientos, representa una cultura, una forma vida y pensamiento). La complejidad de la traducción automatizada debe confrontar y combinar adecuadamente los siguientes factores: las barreras culturales entre lenguajes naturales, la inherente ambigüedad de los lenguajes humanos, la irregularidad entre dos lenguas[22], y su complejidad semántica.

El presente trabajo se enfoca en nuestra investigación de Procesamiento de Lenguaje Natural para el diseño, adaptación y construcción de un sistema de traducción wixarika → español. El wixarika es un idioma que se estima que tiene entre treinta mil y cincuenta mil hablantes[10], con pocos textos escritos, con un análisis gramatical limitado[10,6] y sin un estudio conocido en el campo del Procesamiento de Lenguaje Natural (NLP, del inglés, *Natural Language Processing*). La aplicación del NLP a las lenguas originarias representaría un avance para incorporarlas al nuevo entorno digital.

En México se hablan sesenta y ocho lenguas originarias[3] de las cuales veintiún cuentan con menos de mil hablantes. La UNESCO identificó en el año 2007 que el cincuenta por ciento de las lenguas a nivel mundial se encuentran en peligro de desaparecer, seis mil lenguas son habladas únicamente por el cuatro por ciento de la población mundial y el noventa por ciento de las lenguas no están representadas en Internet[25]. Esto plantea un problema muy importante y trascendente para la cultura universal y los valores humanos, que no es exclusivo de nuestro país: la preservación de la cultura y las lenguas indígenas [12].

El artículo está organizado de la siguiente manera. En la sección 2 se muestran los antecedentes y trabajos previos sobre traducción automática en general y sobre traducción con escasos recursos en particular. La sección 3 describe el modelo de traducción de nuestra propuesta para tratar traducciones de lenguajes polisintéticos y con escasos recursos. Posteriormente se presentan los resultados de las experimentaciones en la sección 4, y por último se presentan conclusiones y trabajos futuros.

2. Teoría del dominio y trabajos previos

Las investigaciones existentes sobre traducción automática para lenguas indígenas son muy escasas, pero si han sido extensamente trabajadas para alemán, español, francés, italiano, portugués, árabe, japonés, koreano, chino, holandés, griego y ruso (en sistemas comerciales y públicos como Google, Systran, Prompt); y en casi todos los casos el inglés es la "contra parte" de las traducciones[16]. Por lo tanto, retomamos las investigaciones del estado de arte y nos centramos en el modelo estadístico por frases, que será el que vamos a modificar para el caso de traducción wixarika → español. Este modelo segmenta la entrada en frases y hace una traducción uno a uno a frases en la lengua objetivo, con un posible reordenamiento [14].

2.1. El idioma wixarika

El wixarika es un idioma perteneciente a la familia yutoazteca , con una estructura sujeto-objeto-verbo (SOV), incorporante y con una fuerte tendencia polisintética, siendo incluso mayor que la del náhuatl. Los morfemas se agrupan en torno a una raíz verbal e incluyen una gran cantidad de información según Iturrio [10]. "La polisíntesis es el resultado de la incorporación de operaciones sintácticas, realizado en otros casos por la combinación de palabras autónomas, a la palabra predicativa, aproximándose al ideal de una palabra por enunciado" [10]. En el siguiente ejemplo se aprecia la forma en que se pueden construir palabras en wixarika a partir de sus reglas silábicas. El concepto de montaña puede ser creado de la siguiente manera.

hai m-a-ta-ka-i-t+ka

Donde *hai* significa nube, y la palabra siguiente es el verbo *matakait+ka* que se divide en morfemas. La combinación entre *m* y *a* refiere a algo figurativo, el *ta* a algo que está al borde de, *ka* localiza esto en cierto espacio, la *i* significa estar, mientras que *t+ka* es plural. El resultado puede ser leído como "donde las montañas bordean", y que de una forma sucinta se traduciría como *montañas*[8]. Es importante destacar, lo que hace el problema de traducción wixarika - español complejo, es que la combinatoria de morfemas se da en torno a la raíz verbal, y no sobre otras palabras.

Al comparar dos frases apareadas es posible observar la distancia entre los dos idiomas a analizar. Tenemos dos frases en español que no varían de manera importante, sin embargo, el cambio en la morfología de su traducción es muy grande.

Quiero quedarme aquí
'ena nep+nehayewakeyu
Quiero que te quedes aquí
ya nep+tinaki'erie 'ena pem+kunauni

Un traductor wixarika - español debe enfrentar estos retos. Una equivalencia palabra a palabra es inoperante, y este es un problema a tratar en nuestra propuesta.

2.2. La traducción automática

Para la tarea de traducción automática (Machine Translation) se han usado varias estrategias, las cuales se pueden dividir en tres grandes campos: la traducción basada en reglas (RBMT del inglés Rule-based machine translation), los modelos estadísticos (SMT, del inglés Statistical Machine Translation) y la traducción basada en ejemplos (EBMT, del inglés Example-Based Machine Translation)[2], además de modelos híbridos que combinan varios aspectos de ellos.

RBMT. En este modelo existen reglas que definen el análisis de los enunciados origen, reglas de cómo transferir las representaciones y finalmente reglas para generar texto de la representación transferida [2]. Este proceso es conocido como análisis–transferencia–generación (ATG). En el caso de que sus reglas sean aplicadas exactamente al caso de traducción, el resultado será de alta calidad y muy preciso, con la ventaja de poder explicar el resultado de la traducción. Pero no es frecuente que sus reglas apliquen a los casos analizados, pues continuamente existen conflictos de reglas o múltiples reglas aplicadas en un mismo caso [2].

SMT. En la traducción máquina estadística las reglas de traducción ATG no son creadas a priori usando los conocimientos lingüísticos, sino que son generados a partir de un conjunto de textos emparejados. Las reglas y palabras son aprendidas de los datos de entrada y son traducidos basados en probabilidades [2]. Estos modelos requieren un gran número de datos para poder funcionar correctamente.

La SMT tiene dos grandes vertientes, la traducción por palabras y la traducción por frases. La primera fue popular en los años ochenta del siglo pasado con el proyecto *Candile* de IBM. La traducción se basa en la probabilidad de que dada una palabra en el origen corresponda a una palabra en el destino. Con una cantidad de datos apareados, esta probabilidad será el número de veces que aparecen las palabras destino cuando aparece la palabra origen en el mismo enunciado emparejado. La segunda, el modelo por frases es el que mejores resultados producía era el estadístico por frases. Este modelo segmenta la entrada en frases y hace una traducción uno a uno a frases en la lengua objetivo con un posible reordenamiento [14]. Los modelos basados en palabras habían demostrado estar limitados por la falta de existencia de una relación uno a uno entre las palabras de dos idiomas. Por ejemplo, una palabra en español no necesariamente corresponde a una en inglés. En ocasiones puede corresponder a una o más palabras, y de igual manera en orden contrario. Esto conlleva además a que un grupo de palabras logren realizar mejor una desambiguación que palabras aisladas. Ahora bien, también es una pregunta importante ¿que se considera como una frase?. El modelo en realidad no tiene conocimiento respecto a esto, aunque un algoritmo complementario puede acotar este tema.

2.3. Traducción con bajos recursos

Para el uso de modelos de STM serían necesarios al menos 100 MB de texto pre-alineado[16], lo cual con idiomas como el wixarika sería imposible de obtener. Para enfrentar este problema se puede recurrir a los modelos RBMT o trabajar en algoritmos híbridos cómo los propuestos por Laukaitis[16], Yaser[1] y Nießen[18]. Asumir una traducción gramatical basada en reglas tampoco es posible por la falta de un cuerpo completo de la gramática wixarika.

Nießn y Ney proponen la utilización de un analizador morfológico que descomponga las palabras en sus raíces

y morfemas para etiquetar posteriormente cada componente. Se auxilia de un diccionario jerárquico que auxiliará a la traducción. Este mecanismo logra reducir el corpus paralelo necesario hasta a 10 % del normalmente necesario. Laukaitis analiza el caso de un traductor asimétrico, donde un lenguaje tiene una gran cantidad de recursos y el segundo carece casi por completo de ellos, con excepción de un analizador morfológico. Con ayuda de un corpus paralelo reducido (de 1 MB) y redes ontológicas del lado del idioma más analizado, logra buenos resultados.

3. Modelo de traducción

Un modelo de un traductor SMT se compone de una fase de entrenamiento que generará un modelo de lenguaje, un modelo de traducción y un modelo de alineamiento. Estos tres modelos estadísticos servirán al decodificador generar posibles traducciones y evaluarlas, intentando encontrar con ello una traducción óptima. Agregamos también una evaluación de la traducción, que permitirá tener una métrica de los resultados del traductor. La fase de entrenamiento y el decodificador serán explicados a continuación.

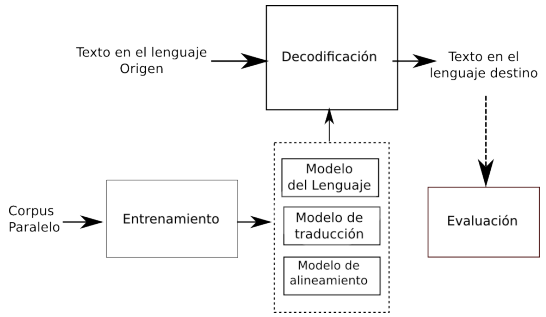

Figura 1. Modelo de un traductor SMT

3.1. Entrenamiento

Sean $f^J = (f_1, \ldots, f_j, \ldots, f_J)$ una frase origen compuesta por una tupla de palabras f_j y $e^I = (e_1, \ldots, e_i, \ldots, e_I)$ una frase objetivo compuesto por palabras e_i, se define un alineamiento $\mathbb{A} \subset \{(i,j) : j = 1, \ldots, J; i = 1, \ldots, I\}$ [20]. Los alineamientos $i = a_j$ pueden contener una palabra vacía e_0. Si se supondría que una palabra f_i tiene una única palabra alineada en e_j o e_0, se obtendría una función de alineamiento $j \to i = a_j$, y no una relación.

Para realizar la traducción se requiere un modelo de alineamiento en su fase inicial de entrenamiento. Si definimos una traducción como probabilidad $p(f'^J|e^I)$ se introduce un factor de alineamiento oculto $p_\theta(f'^J, a^J|e^I)$, siendo el valor de θ un valor desconocido a encontrar. Se define la relación entre la probabilidad de traducción y el modelo de traducción como [27]:

$$p(f^J|e^i) = \sum_{a^J} p_\theta(f^J, a^J|e^I) \tag{1}$$

Sean $S = \{(f_s, e_s) : s = 1, \ldots, S\}$ un conjunto de frases alineadas de un corpus paralelo. Para cada par alienado se encuentra el valor de θ y busca la máxima esperanza, como se menciona en la ecuación 2.

$$\hat{\theta} = \text{argmax}_\theta \prod_{s=1}^{S} \sum_{a} p_\theta(f_s, a|e_s) \tag{2}$$

Para cada enunciado existe una gran variedad de alineamientos $\hat{a}$ estimados, pero se tratará de encontrar el mejor alineamiento, también llamado alineamiento Viterbi, tal que

$$\hat{a}^J = \text{argmax}_{a^J} p_\theta(f^J, a^J|e^I) \tag{3}$$

Dado que el wixarika es un lenguaje polisintético, el alineamiento con palabras al español es poco prometedor. Los afijos se aglutinan en torno el verbo, tanto antes de la raíz verbal como después. La función $j \to i = a_j$ no se cumple como un mapeo uno a uno, sino en forma de relación $j \to (i_1 \ldots i_n) = a_j$ donde $k \geq 1$, y a_j es una tupla de pares de alineamiento. Se crea una función Γ, que descompone las palabras $f_j^I \to M_j$ en una lista de morfemas ordenados $(m_1, \ldots, m_n, \ldots, m_N)$, donde N es el número total de morfemas. El nuevo conjunto M'^K será:

$$f'^K = \bigcup_{j=1}^{J} \Gamma(M_j) \tag{4}$$

La cardinalidad de $K = |f'|$ es el número de todos los morfemas y palabras generados evaluando en Γ todas las palabras de la frase original. La tupla de frases f^J se sustituye por la nueva tupla F'^K en la ecuación de alineamiento 2 y en el modelo de traducción estadístico en la ecuación 7. La figura 2 muestra la mejora en el alineamiento de palabras del modelo de descomposición morfológica al modelo de alineamiento de palabras. La frase *"ik+ ki p+kahekwa"* se traduce como *"esta casa no es nueva"*. Pero la palabra *"p+kahekwa"* contiene la información de tres palabras en español. Si usamos nuestra función Γ(p+kajekwa) obtendríamos la tupla $(p+, ka, hekwa)$. La unión de todas las palabras descompuestas y no descompuestas de la frase original f^I, nos genera un mejor alineamiento respecto al español. La función Γ es un Trasductor de Estados Finitos, con la información morfológica descrita en [10] y [8]. Los idiomas polisintéticos y aglutinantes comparten la característica de poder ser expresados mediante un trasductor, como es el caso del turco [5] [4].

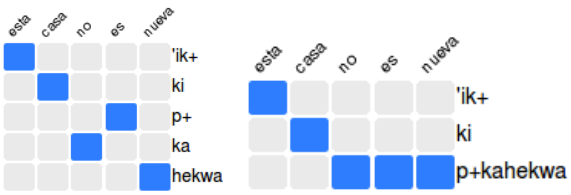

Figura 2. Búsqueda de la mejor traducción

Para el caso de idiomas con gran riqueza morfológica como el caso de estudio se sugiere separar los morfemas que sean más parecidos a palabras del inglés, conservar unidos los morfemas (como tiempos verbales) a sus raíces que se comporten de manera semejante en inglés e ignorar los que no tienen funciones parecidas [14]. En nuestra aplicación se va a tomar la integridad de los morfemas de f' para conservar la mayor cantidad de información posible.

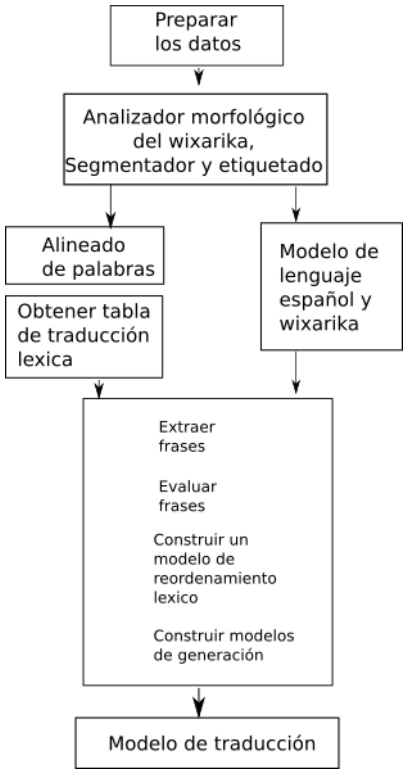

Figura 3. Proceso de entrenamiento

El modelo de alineación es usado tanto por el modelo de traducción por palabras y el de frases. Pero a diferencia de un modelo por palabras, donde se escoge la mejor alineación, palabra por palabra, en el modelo por frases se escoge un número de pares de frases y se evalúa en $\text{count}(\bar{e}, \bar{f})$. La probabilidad de traducción es estimada en una frecuencia relativa que será nuestro modelo ϕ [14]:

$$\phi(\bar{f}'^J|\bar{e}^I) = \frac{\text{count}(\bar{e}^I|\bar{f}'^J)}{\sum_{\bar{f}_i'^J} \text{count}(\bar{e}^I, \bar{f}_i'^J)} \tag{5}$$

El proceso de entrenamiento, como se ilustra en la figura 3.1, requiere la preparación de los datos, la segmentación morfológica (ecuación 4), un alineamiento de palabras (ecucación 1), el entrenamiento de un modelo de lenguaje(comunmente usando n-gramas) y la generación de un modelo de traducción denominado ϕ (ecuación 5). En el entrenamiento se requiere un algoritmo de extracción de frases y el cálculo de una tabla de probabilidades de traducción. En la figura 3 se presenta el proceso de entrenamiento con una nueva etapa de análisis morfológico, descomposición y etiquetado. En lo que a la decodificación se refiere se deben insertar dos nuevas etapas, una descomposición morfológica antes de la codificación.

3.2. Traducción

Si bien el modelo de alineamiento había sido creado para los modelos basados en palabras, estos habían demostrado estar limitados por la falta de existencia de una relación uno a uno entre las palabras de dos idiomas, problema que persiste a pesar de la descomposición morfológica que proponemos en la ecuación 4. Esto conlleva además a que un grupo de palabras logren realizar mejor una desambiguación que palabras aisladas. En la definición del modelo matemático según Kohen[14] se usa la regla de bayes para invertir la dirección de traducción e integrar un modelo de lenguaje que se define como p_{LM}.

$$e_{mejor} = \text{argmin}_e^I p(e^I|f'^J) \tag{6}$$

$$= \text{argmin}_e^I p(f'^J|g^I) p_{LM}(e) \tag{7}$$

Para el modelo por frases se va a descomponer $p(f'^J|e^I)$ en:

$$p(\hat{g}_1^I|\hat{s}_1^I) = \prod_{i=1}^{I} \phi(\hat{g}_i|\hat{s}_i) d(\text{inicio}_i - \text{fin}_{i-1} - 1) \tag{8}$$

Cada una de las frases origen f es segmentada en I frases $\bar{f}_i$ y como se comentó, por el teorema de bayes se invierte la probabilidad para modelar la traducción e^I a f^J a través de un canal ruidoso $\phi(\hat{g}_i|\hat{s}_i)$. La distancia mide el inicio de la frase origen al fin de la misma, y es el número de palabras que se van a omitir cuando se toman las frases origen fuera de su enunciado[14].

En la figura 4 se muestra al flijo de una cadena entrante $\bar{f}^J$. El texto entrante necesita ser preparado mediante una normalización y un tokenizado. Una vez preparado se procede a un análisis mofológico y a su segmentación y etiquetado.

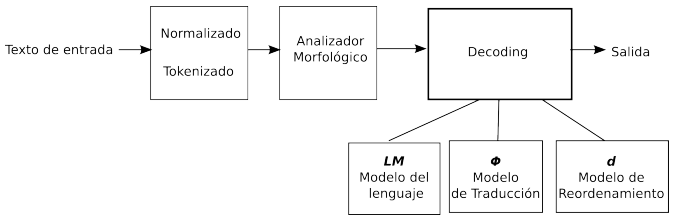

Figura 4. Flujo de traducción

Con el modelo de lenguaje LM, el modelo de alineamiento d y el modelo de traducción ϕ se busca la traducción con el mejor puntaje en el modelo expresado en la fórmula 2. Al ser este problema de combinatoria un problema NP-Completo [13], se requiere el uso de heurísticos para encontrar una traducción aproximada. Se utilizan algoritmos como Beam o A* [9] [11] para ese fin. La búsqueda resultante es expresada en grafos, como se precia en la figura 5.

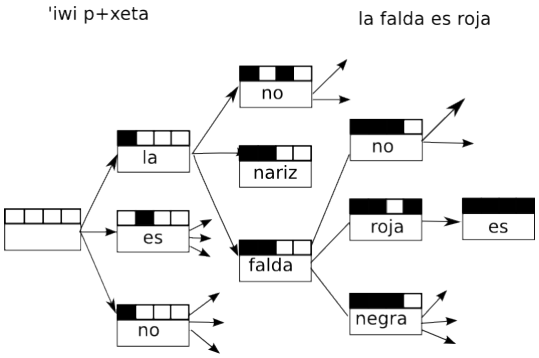

Figura 5. Búsqueda en el espacio de traducción

4. Experimentos y Resultados

Las pruebas se realizaron en una computadora con dos procesadores Intel Xeon X3450 x86 de 64 bits con 4 núcleos cada uno y capacidad de dos hilos por núcleo, a 2.67 GHz NUMA, con 16 GB de memoria RAM. Para el sistema de alineado usamos GIZA++ [19] y como sistema de decodificación se utiliza el sistema MOSES [15] usando el modelo de entrenamiento de traducción por frases.

El corpus usado fue extraído del libro [8] que aporta valiosa información morfológica en su texto apareado. Se utilizaron 100 frases apareadas como corpus de experimentación, y las traducciones se realizaron únicamente con los afijos y las palabras usadas en el corpus. Para la evaluación no se utilizó BLEU [21] por el reducido corpus, y las frases de tamaño variable según el modelo de descomposición morfológica, y se prefirió WER [26] y TER [23], que son eficientes en estas condiciones. Con los valores obtenidos se realiza una comparación de resultados (ver tabla 1), entre una traducción de alineamiento por palabra (ecuación 2) en comparación con la descomposición morfológica descrita en la ecuación 4.

	WER	TER
Sin segmentación morfológica(SGM)	38	0.84
Con segmentación morfológica(CSM)	25	0.46
Segmentación con etiquetado(CSEM)	21	0.46

Cuadro 1. Evaluación de traducción

El error en la traducción automática usando palabras sin segmentación es mas alto que si usamos un segmentador morfológico. Los resultados usando además un etiquetador de morfemas son ligeramente superiores al hecho de no usarlo.

La técnica con segmentación y etiquetado tiene una clara ventaja con respecto a una segmentación simple y la traducción sin segmentación. Para ilustrar la diferencia en la calidad de traducción, se muestra una tabla comparativa 2.

Wixarika	Sin Segmentar	Segmentado
neki	neki	mi casa
'aki p+tuxa	'aki es blanca	tu casa blanca
hakewa ne ki	esta falta es no es nueva	esta falda no es nueva

Cuadro 2. Ejemplos de traducción

5. Conclusiones y trabajos futuros

La mayor parte de lenguas originarias del continente americano, incluido el quechua y el aimara, son aglutinantes, y por lo tanto con una gran complejidad morfológica. Para estos idiomas, no es posible retomar a integridad el modelo de alineación por palabras y traducción por frases. Los resultados obtenidos con segmentación morfológica son significativamente mejores que sin usarla. La limitante del corpus paralelo, también, es mejorada mediante la segmentación morfológica.

Para trabajos futuros, estamos valorando posibles mejoras como la ampliación del corpus mediante diccionarios o con técnicas de extracción de corpus paralelo, como se ha estudiado para el Nahuatl por Gutiérrez [7].

Existen además traductores estadísticos de idiomas indígenas, con una implementación cerrada, desarrollados por *Microsoft Translator Community Partners*[17], para el otomoí de Querétaro y el maya de Yucatán. Nuestra postura es proveer una herramienta de software libre wixarika-español para los fines que las personas de los pueblos y comunidades requieran. Por ejemplo, para libremente seleccionar que libros y textos traducir.

Nuestra investigación, hasta donde conocemos, es la primer aplicación de NLP al wixarika, con sus trabajos futuros permitirá avanzar en otras lenguas indígenas y actuar como una Piedra Rosetta de nuestros tiempos. Para trabajos futuros es posible incorporar interfaces de voz y tinta electrónica, que facilitarán la interacción de las personas con sistemas de traducción automática. El progreso en la tecnología de tabletas, celulares y procesadores programables hacen atractivo el diseño y construcción de una aplicación o de un dispositivo de traducción automática personal. Este tipo de herramientas fomentan la vitalización de las lenguas originarias en un entorno marcado por las TIC.

Referencias

1. Yaser Al-Onaizan, Ulrich Germann, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Daniel Marcu, and Kenji Yamada. Translation with scarce bilingual resources. *Machine Translation*, 17(1):1–17, 2002.
2. Pushpak Bahattacharyya. *Machine Translation.* CRC Press, 2015.
3. Instituto Nacional de Lenguas Indígenas. Lenguas indígenas en méxico y hablantes (de 3 años y más) al 2015, January 2016.
4. Marina Ermolaeva. An adaptable morphological parser for agglutinative languages. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 164–168. Pisa University Press, 2014.
5. Gülşen Eryiğit and Eşref Adalı. An affix stripping morphological analyzer for Turkish. In *Proceedings of the International Conference on Artificial Intelligence and Applications*, pages 299–304, Innsbruck, 16-18 February 2004.
6. Joseph E. Grimes. *Huichol Sintax.* Series Practica. Mouton & Co, 1964.
7. Ximena Gutierrez-Vasques. Bilingual lexicon extraction for a distant language pair using a small parallel corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 154–160, Denver, Colorado, June 2015. Association for Computational Linguistics.
8. Paula Gómez. *Huichol de San Andrés Cohamiata, Jalisco.* Archivo de lenguas indígenas de México. Colegio de México, 1999.
9. P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems, Science, and Cybernetics*, SSC-4(2):100–107, 1968.
10. José Luis Iturrio and Paula Gómez López. *Gramática Wixarika I.* Archivo de lenguas indígenas de México. Lincom Europa, 1999.
11. Mager Hois Jesús Manuel. El algoritmo fringe search como solución superior a a* en la búsqueda de caminos sobre gráficos de malla, May 2015.
12. Mager Hois Jesús Manuel. Traductor wixárika-español, May 2016.
13. Kevin Knight. Decoding complexity in word-replacement translation models. *Comput. Linguist.*, 25(4):607–615, December 1999.
14. Philipp Koehn. *Stadistical Machine Tanslation.* Cambridge University Press, 2010.
15. Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
16. Algirdas Laukaitis and Olegas Vasilecas. *Computational Linguistics and Intelligent Text Processing: 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18-24, 2007. Proceedings*, chapter Asymmetric Hybrid Machine Translation for Languages with Scarce Resources, pages 397–408. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.
17. Microsoft. Microsoft translator community partners, 3 2016.
18. Sonja Nießen and Hermann Ney. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 20(2):181–204, June 2004.
19. Franz Josef Och. An efficient method for determining bilingual word classes. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 71–76, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.
20. Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.
21. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
22. Maxim Roy. *Approaches to handle scarce resources for Bengali Statistical Machine Translation.* PhD thesis, Simon Fraser University, Burnaby, BC, Canada, 4 2010.
23. Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006.
24. Alfred Tarski. Der wahrheitsbegriff in den formalisierten sprachen. *Studia Philosophica*, 1:261–405, 1936.
25. UNESCO. ElaboraciÓn de una convenciÓn para la protecciÓn de las lenguas indÍgenas y las lenguas en peligro, 4 2007.
26. Klaus Zechner and Alex Waibel. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 186–193, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.
27. Richard Zens, Franz Josef Och, and Hermann Ney. *Phrase-Based Statistical Machine Translation*, pages 18–32. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.