EVALUACION HUMANA DE SISTEMAS DE T.A

PROYECTO FINAL DE CARRERA

Autor: Ginés Mendoza Mompeán
Director: Fernando Daniel Quesada Pereira

ABSTRACTO

En este diploma hemos desarrollado una herramienta en la cual nosotros podremos evaluar la calidad de sistemas de traducción automática, comparando la traducción de un sistema de traducción con la traducción de un hablante nativo. Esta aplicación nos permite recoger juicios humanos sobre el resultado de la traducción automática. Para que esto sea posible hemos desarrollado algunas herramientas tales como: 1) score the translation, 2) error classification, 3) fluency & adequacy of the translation, 4) mistranslated sentences. Tenemos que decir que la aplicación ha sido desarrollada para mejorar es sistema de traducción automático, aunque existen métodos automáticos para reducir el costes y tiempos de ejecución en el proceso de evaluación. La Evaluación humana sigue siendo necesaria en la investigación de traducción automática (TA) ya que los resultados que se obtienen en una evaluación automática no son del todo exactos.

OBJETIVO

El objetivo de esta aplicación es evaluar la calidad de los sistemas de traducción automática para dar algunas ideas de como los sistemas de traducción automática podrían mejorar. Existen métodos de evaluación automática tales como WER, Meteor o BLUE que miden la calidad de los sistemas de traducción de una forma automática (usando métodos estadísticos). El principal problema de la evaluación automática es como obtener un alto nivel de precisión en las medidas obtenidas de forma automática con respecto a las medidas obtenidas manualmente por un anotador. Usando mediciones automáticas ahorramos coste y tiempo, pero la evaluación manual sigue siendo aún necesaria ya que obtenemos un alto nivel de exactitud y una evaluación detallada de las traducciones de salida. Sin embargo la evaluación humana en los sistemas de traducción automática consume tiempo y costes. Para reducir estos problemas, hemos desarrollado una interfaz gráfica (GUI) que permite al anotador realizar sus evaluaciones de una manera rápida y sencilla.

El desarrollo de una aplicación que recoge juicios humanos de sistemas de traducción automática puede ser una labor complicada ya que el anotador tiene que tener en cuenta las siguientes tareas: seleccionar la puntuación de la traducción, clasificar los errores de las traducciones, juzgar la fluidez y la calidad de la traducción y seleccionar las sentencias mal traducidas. Cada anotador puede tener una manera diferente de interpretar la sentencia y por esta razón la aplicación tiene que tener un diseño que ayude a evaluar las sentencias a traducir.

En este documente describiremos como usar dicha aplicación y como usar las herramientas que nos permitan recoger los juicios humanos de las traducciones de salida.

TRADUCCIÓN AUTOMÁTICA

La traducción automática está dentro del campo de la Lingüística Computacional y usa conocimientos de otros campos, tales como: informática, Lingüística, negocios, etc…

A partir de los 50's y principios de los 60's del siglo XX, Hubo algunos ingenieros americanos especializados en la inteligencia artificial que creían en la posibilidad de traducir textos de forma automática y que habría una posibilidad de que las maquinas llegaran a ser capaz de hacerlo. La traducción automática (TA) empezó como un estudio que podría ser útil para reducir costes de traducción en las empresas y organizaciones internacionales.

Los sistemas TA son capaces de traducir grandes cuerpos de texto en un periodo más corto de lo que un humano puede ser capaz. Proyectos tales como "la traducción automática de páginas web" sería imposible sin la ayuda de sistemas de traducción automática. Por otro lado, La TA también se convirtió en una necesidad para organizaciones internacionales tales como la Comunidad Europea, la cual tiene que generar una gran cantidad de documentos en diferentes lenguas y en un tiempo ilimitado. Por esta razón, la comunidad financio el proyecto "Eurotrans", con el objetivo de desarrollar un sistema capaz de traducir los documentos de forma automática en todas las lenguas de la Unión Europea.

La TA es más exitosa cuando traduce textos escritos en un lenguaje controlado. Un documento está escrito en un lenguaje controlado, si este tiene una estructura sintáctica, no es ambiguo y tiene un vocabulario limitado.

LIMITACIONES

Las limitaciones de un sistema de TA altera la calidad de la traducción. Si un sistema de TA no tiene una presentación apropiada del significado de la frase de origen, es probable que la traducción tenga una mala traducción o sea ilógica.

La compresión de una frase requiere conocimiento complejo de la lengua de origen y algunos elementos para poder procesar la información lingüística. Obviamente, el procedimiento de todo esto costaría mucho esfuerzo y conllevaría mucho tiempo y probablemente los recursos de memoria del sistema podrían venirse abajo bruscamente.

Hoy en día, existe un alto nivel de calidad de traducción entre lenguas romances (Español, Portugués, Catalán, etc…). Sin embargo, los resultados empeoran cuando las lenguas no son similares entre ellas, como es el caso del español con el inglés o el alemán.

Otro punto que influye en la calidad de la traducción es el grado de especialización en el sistema de traducción. La calidad de la traducción puede ser mejorada si el sistema de traducción está especializado en un tipo de texto y en un vocabulario especifico. Por ejemplo, un sistema especializado en la traducción de documentos sobre el tiempo, tendrá un alto nivel de calidad incluso traduciendo textos entre lenguas totalmente diferentes, pero este sistema será inútil para otro tipo de documentos tales como: deportes, financiación, automovilismo, etc…

La traducción es una tarea dura que requiere mucho conocimiento y habilidad. En una traducción no es suficiente cambiar una palabra por otra, sino que también hay que ser capaz de reconocer todas las palabras en el contexto y la influencia que ellas tienen una sobre la otra. El lenguaje humano tiene una morfología específica, sintáctica y semántica. Por lo que un texto aparentemente simple puede llegar a ser muy ambiguo. Es necesario tener en cuenta las cuestiones de estilo, discurso y pragmática.

Sin embargo, Existen métodos estadísticos que realizan traducciones sin tener en cuenta cuestiones gramaticales. Hoy en día la tendencia es integrar todo tipo de metodologías: conocimiento explícito de la lengua y estadísticos de un corpus.

EVALUACIÓN

El estudio de máquinas de traducción automática necesita un juicio apropiado de las traducciones obtenidas, consistente y fácil de usar para la evaluación de los resultados. Es necesario tener algunos mecanismos que puedan comparar diferentes sistemas entre sí, o averiguar cómo afecta cualquier variación del sistema de traducción automática a la calidad de las traducciones.

La evaluación del sistema de traducción presenta una serie de dificultades. Como por ejemplo, la dificultad que resulta definir un proceso subjetivo. Frecuentemente encontramos diferentes enfoques en el mundo de la traducción automática.

La calidad de una traducción puede ser expresada en dos atributos principales: la fidelidad y la fluidez del texto. Mientras que la fluidez es una evaluación monolingüe la fidelidad es una evaluación bilingüe y por lo tanto, es un proceso más costoso.

MEDIDAS "DARPA" DE EVALUACIÓN PARA LA TRADUCCIÓN

Del 1992 hasta el 1994, DARPA promovió una serie de iniciativas para definir medidas para evaluar los sistemas de traducción automática. Como resultado de tales proyectos, surgieron tres tipos de medidas:

Adecuación:

Los evaluadores miden la exactitud del significado en la salida del sistema de traducción comparándolo con el significado de la traducción de referencia. Por esta razón, mostramos una traducción de referencia creada por un experto, junto con la traducción creada por un sistema de traducción automática. El evaluador dispondrá de una escala del 1 a N para evaluar la salida. Es por lo tanto una medida de fidelidad.

Información transferencia:

El evaluador responde preguntas con múltiples respuestas sobre el texto traducido, como si de un comentario de texto se tratara. Es otra medida de fidelidad.

Fluidez:

La Fluidez mide la calidad de una traducción en acuerdo con el grado de exactitud con respecto al lenguaje destino, sin tener en cuenta la sentencia de origen. Los evaluadores pueden ver las sentencias propuestas, ellos deberán evaluarlas del 1 a N, en función de cómo fue aceptada intuitivamente por el hablante nativo, además de que tendrán que considerar la corrección gramatical y si el corpus del texto traducido tiene sentido con el contexto del conjunto del texto traducido.

EVALUACION AUTOMÁTICA

En la evaluación automática es común el uso de medidas objetivas que pueden ser evaluadas automáticamente. Estas medidas toman como referencia una posible traducción de referencia para cada una de las sentencias que queremos traducir. Esta referencia será comparada con la sentencia propuesta por el sistema de traducción. Las medidas más importantes son: Word Error Rate (WER), Sentence Error Rate (SER), Position-Independent WER (PER), Multi reference WER (mWER), Bilingual Evaluation Understudy (BLUE).

EVALUACION HUMANA

En la evaluación Humana es imprescindible la presencia de una persona para obtener la evaluación de la traducción del sistema de TA. Entre las medidas más usadas, podemos destacar las siguientes:

Subjetive Sentence Error Rate (SSER, porcentaje subjetivo de frases erróneas):

Cada sentencia es puntuada del 0 al N, con respecto a la calidad de la traducción. El mayor problema que esta medida presenta es la subjetividad, ya que dos perdonas pueden tener diferentes criterios para evaluar la misma sentencia. Otra desventaja es que no se tiene en cuenta la longitud de la frase. La puntuación de una sentencia de 50 palabras tiene el mismo impacto que una sentencia con solo dos palabras.

Information Item Error Rate (IER, porcentaje de ítems informativos erróneos):

Está medida intenta responder a la siguiente cuestión: ¿Qué debemos hacer si en una frase larga y hay partes de la frase con una buena traducción y otras partes de la frase con una mala traducción? Para resolver este problema, introducimos el concepto “ítems informativos”. Cada ítem de la frase de entrada es calificado como “ok”, “fail”, “syntactic” o “others, dependiendo del resultado en la traducción. La Medida IER puede ser calculada como el porcentaje de ítems mal traducidos.

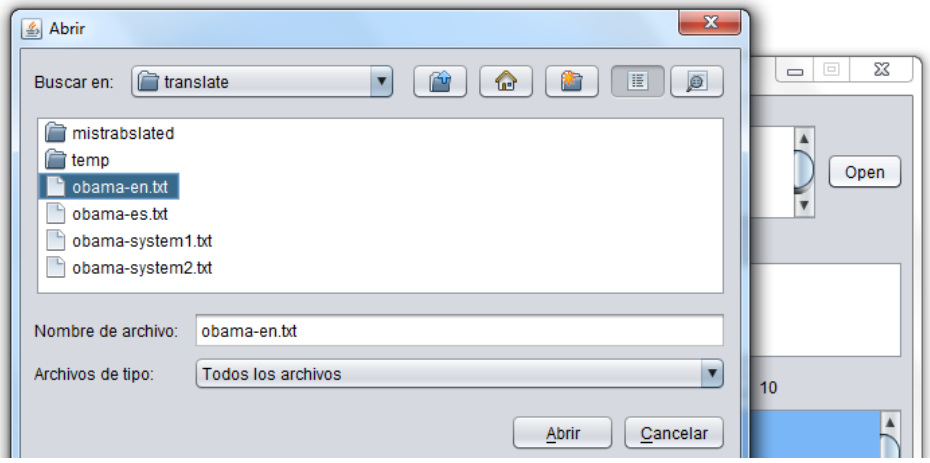
Information Item Semantic Error Rate (ISER, porcentaje de ítems informativos semánticos erróneos):

Esta medida es una modificación de IER, donde un ítem es considerado correcto, si la información deseada es transmitida, sin tener en cuenta errores sintácticos.

HERRAMIENTAS PARA LA EVALUACION HUMANA DE TA

Como abrir los ficheros.

El anotador tendrá que seleccionar la carpeta de referencia (pulsando en el botón “open”), donde estarán almacenadas las frases de origen (frases que tendremos que evaluar) junto a las posibles traducciones de referencia (frases que han sido traducidas por el sistema TA).



Nosotros solo tendremos que seleccionar el archivo donde se encuentran almacenadas las frases a traducir (en nuestro caso “obama-en.txt”), los demás archivos (archivos que contienen las frases traducidas por el sistema de TA) serán abiertos de forma automática.

PARTES DE LA APLICACIÓN

La aplicación puede ser dividida en cinco partes:

Frames:

En esta parte de la aplicación podemos distinguir cuatro cuadros (frames) donde podemos diferenciar las traducciones de los diferentes sistemas de TA junto a la traducción hecha por el hablante nativo y la frase de origen.

Puntuación de la traducción:

En esta parte hemos utilizado el método “Subjective Sentence Error Rate” para evaluar la traducción de la TA. El método SSER fue descrito en el apartado de “Evaluación Humana”

Hemos usado una puntuación del 1 al 10, siendo 1 la peor puntuación y 10 la mejor.

Clasificación de errores:

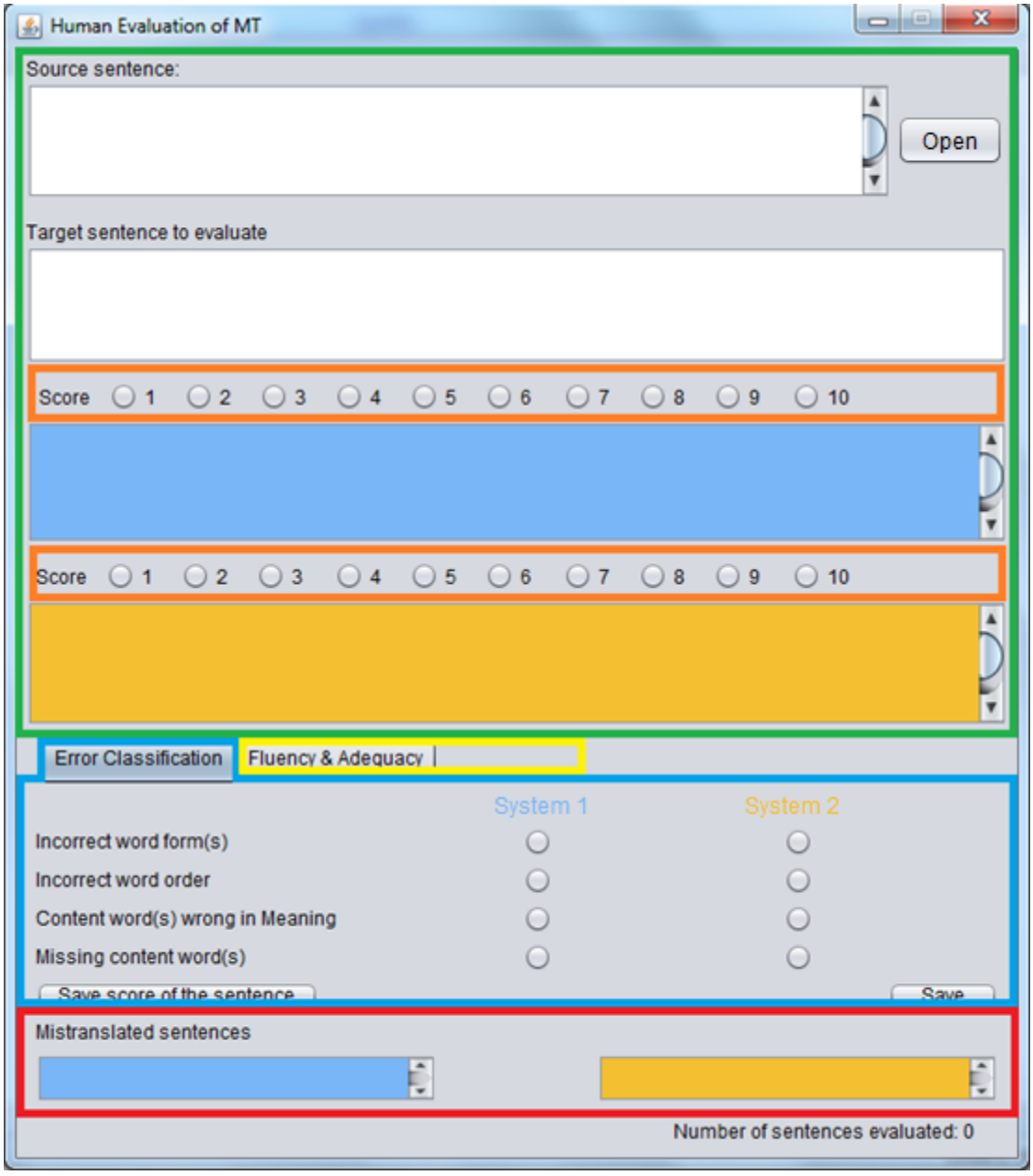
En esta parte hemos desarrollado una herramienta para analizar los errores que la traducción pueda tener. Hemos creado una sección donde el anotador tiene que clasificar los diferentes errores que la traducción pueda tener. Los erros que hemos considerado más importantes para la traducción son: forma incorrecta la palabra/s, orden de las palabras incorrecto, palabra/s con un significado incorrecto y falta de contenido en la traducción.

Fluidez y Adecuación:

Esta herramienta nos permite medir la fluidez y adecuación de la frase traducida, el anotador tendrá que evaluar la fluidez y la adecuación de la traducción teniendo en cuenta la traducción hecha por el hablante nativo.

Frases mal traducidas:

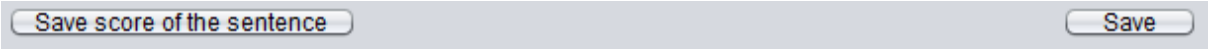
En esta parte de la aplicación el anotador tendrá que seleccionar la parte de la frase con una mala traducción producida por el sistema de TA. Una vez hecha la selección, El anotador tendrá que escribir la traducción que el considere correcta.



COMO GUARDAR LOS RESULTADOS DE LA EVALUACIONES HECHAS POR EL ANOTADOR

Una vez que el anotador ha abierto el archivo “Obama-en.txt” y las frases están lista para evaluar, el anotador debe iniciar el proceso de evaluación. Cuando el anotador ha hecho la evaluación de la primera traducción deberá pulsar el botón “Save score of the sentence” (todas las puntuaciones serán refrescadas de forma automática para evitar confusiones) y la aplicación de forma automática pasara a la siguiente traducción a evaluar. Cuando el anotador ha terminado de valorar todas las traducciones, los cuadros estarán en blanco, cuando esto suceda el anotador tendrá que pulsar el botón “Save”. Con este botón la aplicación generara dos archivos (.txt) donde se encontrarán todos los resultados y estadísticas de la evaluación humana de cada sistema de TA.

Los archivos temporales estarán guardados en el directorio “…\NetBeansProjects\ManualEvaluation\translate\temp”, mientras que los archivos “mistranslated” estarán guardados en el directorio “…\NetBeansProjects\ManualEvaluation\translate\mistrabslated”.



RESULTADOS

Como hemos mencionado anteriormente, cuando el anotados guarda todas las evaluaciones que ha hecho de las traducciones, La aplicación genera un documentos con los resultados de la evaluación humana para cada sistema de TA. A continuación explicaremos los resultados:

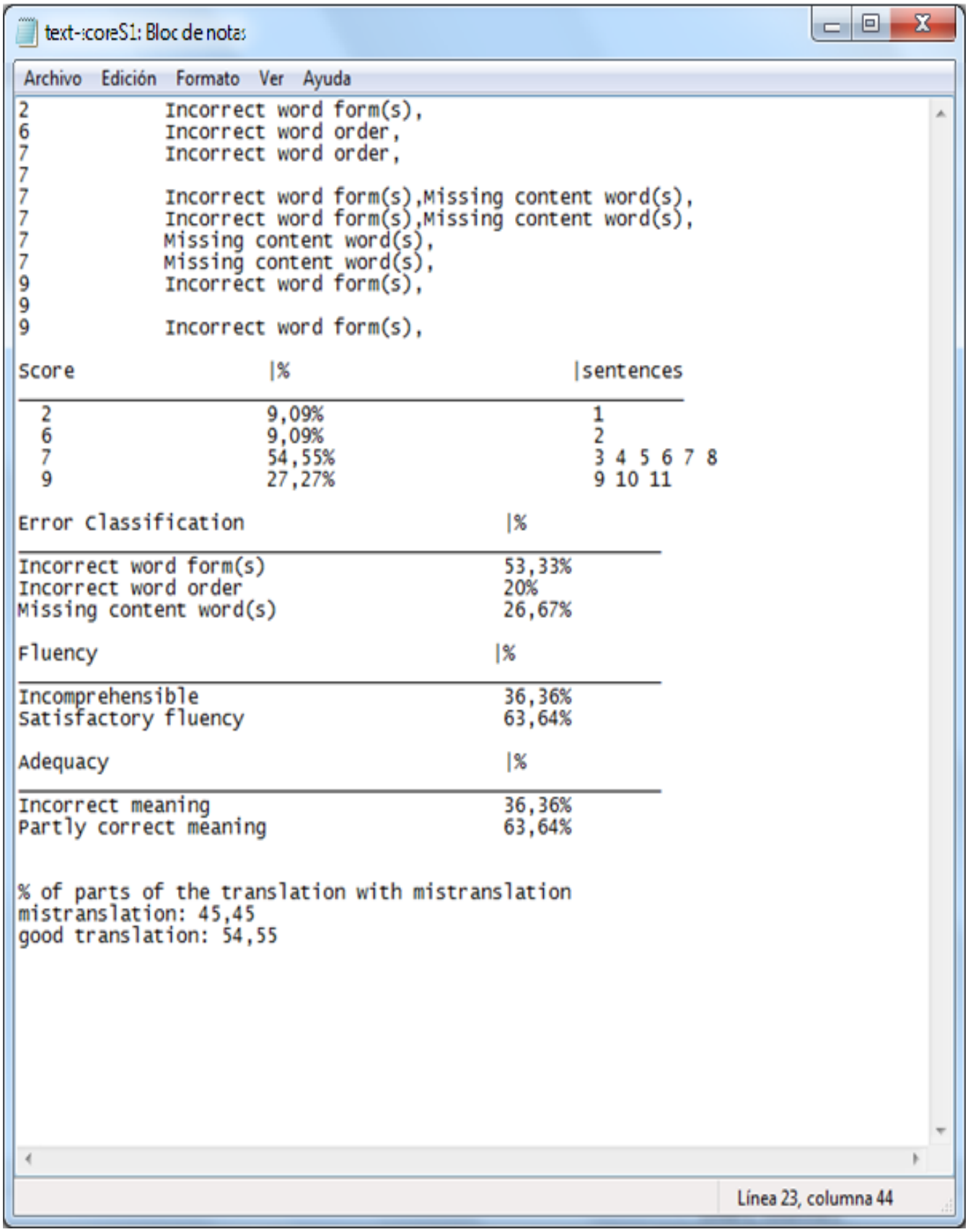
En dicho documento podemos distinguir cuatro grandes secciones:

La primare sección del documento está separada en dos columnas. En la primera columna se encontraran las puntuaciones de cada frase, mientras que en la segunda columna podremos ver los erros de clasificación que la frase pueda tener. La puntuación y los errores de clasificación de la primera frase estarán situados en la primera línea del documento, la puntuación y los errores de clasificación de la segunda frase estada situado en la segunda línea y así sucesivamente.

La segunda gran sección podremos encontrar la puntuación de cada sentencia y el porcentaje de cada puntuación.

La tercera sección del documento estará los porcentajes de error de clasificación, fluidez y adecuación.

Finalmente, en la cuarta sección tendremos el porcentaje de las frases que han tienen parte de la frase mal traducida.


text-scoreS1: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
2             Incorrect word form(s),
6             Incorrect word order,
7             Incorrect word order,
7
7             Incorrect word form(s),Missing content word(s),
7             Incorrect word form(s),Missing content word(s),
7             Missing content word(s),
7             Missing content word(s),
9             Incorrect word form(s),
9
9             Incorrect word form(s),

Score                   |%                          |sentences
------------------------------------------------------------
  2                     9,09%                       1
  6                     9,09%                       2
  7                     54,55%                      3 4 5 6 7 8
  9                     27,27%                      9 10 11

Error Classification                    |%
------------------------------------------------------------
Incorrect word form(s)                  53,33%
Incorrect word order                    20%
Missing content word(s)                 26,67%

Fluency                                 |%
------------------------------------------------------------
Incomprehensible                        36,36%
Satisfactory fluency                    63,64%

Adequacy                                |%
------------------------------------------------------------
Incorrect meaning                       36,36%
Partly correct meaning                  63,64%


% of parts of the translation with mistranslation
mistranslation: 45,45
good translation: 54,55
```

Línea 23, columna 44

HERRAMIENTA DE DESARROLLO

Java es un lenguaje de programación de alto nivel. Es un lenguaje de propósito general, concurrente, orientado a objetos y basado en clase. Las aplicaciones que son creadas con Java no dependen del hardware en el cual están corriendo, lo que permite ser programadas una vez y ser ejecutado en diferentes sitios.

Esta característica hace que Java sea apropiado para corporativas y aplicaciones web, donde podemos encontrar diferentes plataformas: Windows, Linux, Unix, Mac, etc. Cuando nosotros compilamos un programa en Java, inmediatamente es generado un código en el mismo lugar donde el código fue generado. Este código es conocido como bytecode y es interpretado en el ordenador en el cual está corriendo.

Para que esto sea posible, es necesario que el ordenador pueda interpretar el código en el cual el bytecode está corriendo. Por esta razón, los ordenadores tienen que tener lo que nosotros conocemos como máquina virtual de Java (JVM). La máquina virtual de Java no está implementado por defecto en los ordenadores, por lo que tendremos que instalarla, una de las maneras de hacer esto, es visitando la página oficial de Oracle y de forma gratuita podremos descargar la máquina virtual de Java.

Sun Microsystem divide Java en tres grandes ramas, cada una de ellas con sus respectivos set de APIs y sus propias herramientas de desarrollo: grandes ordenadores, equipos de escritorio y microordenadores o dispositivas de memoria limitada

Java SE es la versión para equipos de sobremesa , también será la plataforma que usaremos para desarrollar nuestra aplicación.

ENTORNO DE DESARROLLO

La tecnología Java está estrechamente relacionada con el mundo del "Open Source" y esta es una de las ventajas que Java presenta. Por esta razón, es más fácil encontrar una gran cantidad de IDEs gratis. Una de ellas es NetBeans, esta IDE es ofrecida por Sun y es la que utilizaremos para la creación de nuestra aplicación.