Wir reden heute über die Qualitätssicherung der maschinellen Übersetzung. Qualitätssicherung bedeutet, dass ein Text grundsätzlich zwar maschinell übersetzte wird, jedoch weitere Maßnahmen getroffen werden, um die resultierende Übersetzung zu verbessern oder zu evaluieren wie effektiv ein Übersetzungsprogramm ist.
Die Qualitätssicherung kann, auch wenn eine Maschine den Text übersetzt hat, einiges an Zeit in Anspruch nehmen, ist jedoch sehr oft trotzdem weniger Arbeit als den Text selbst zu übersetzen.
Bei der Qualitätssicherung gibt es zwei Ansätze: die menschengestützte und maschinengestützte Qualitätssicherung, welche für unterschiedliche Zwecke verwendet werden.
Bei der maschinengestützen Qualitätssicherung geht es weniger um das Korrigieren des Textes, sondern mehr um das Feststellen der Effizienz des Übersetzungsprogramms indem es von Menschen übersetzte Referenztexte mit einem maschinell Übersetzten Text vergleicht.
Diese Programme sind dadurch auch nicht in der Lage die Sinnhaftigkeit des Textes zu überprüfen.

Der Großteil aller gängigen maschinellen Qualitätssicherungsverfahren basieren auf statistischer Auswertung von Referenztexten, welche dann mit dem Kandidatentext verglichen werden um die Güte des Textes zu erfinden.
Bevor man aber über die Verfahren selbst sprechen kann, muss man ein paar Begriffe definieren. Zuerst ein N-Gramm, welches eine Zerlegung eines Textes in Bausteine ist, wonach diese statistisch ausgewertet werden.
Die genau Natur dieser Bausteine ist abhängig von dem Feld, in welches es eingesetzt wird. In der Chemie ist zum Beispiel ein Element ein Baustein, während es in der Linguistik entweder ein Buchstabe oder ein Wort ist.
Das N in N-Gramm steht für die variable Länge dieser Bausteine.
Es gibt Monogramme, welche nur aus einem Zeichen bestehen, Bigramme, welche aus zwei Zeichen bestehen und so weiter mit Tri- und Tetragrammen. Abhängig von der Länge der Gramme kann man dann andere Sachen über den Text erfahren.
Bei einem Monogram erfährt man zum Beispiel die Anzahl der Vorkommnisse der Wörter in dem Text.
Bei Bigrammen erfährt man welche Bausteine auf bestimmte andere Bausteine folgen.
Das ist eventuell mit einem Beispiel ersichtlicher.
Wenn der Satz "Welcome to come" mit Bigrammen analysiert wird, erhält man diese Liste an Paaren.
Der Unterstrich steht für den Beginn, das Ende oder ein Leerzeichen innerhalb des Textes.
Wenn wir uns die drei fett gedruckten Teile ansehen, sehen wir, dass o folgend auf c, m folgend auf o und e folgend auf m jeweils zwei Mal vorkommen.
Aus diesem Grund würde man, wenn man nur diesen Satz als Referenz hätte, annehmen, dass es doppelt so wahrscheinlich ist, dass ein m auf ein o folgt als, dass ein Leerzeichen auf ein o folgt.
Ein Textkomplettierungsprogramm verwendet zum Beispiel solche Statistiken um eine Vorhersage zu treffen welches Wort man gerade schreiben will. Wie das für die Qualitätssicherung relevant ist wird sich gleich zeigen.
Aber zuerst noch die restlichen Begriffe.
Die Trefferquote, im Englischen recall genannt, zeigt wie viele der aufzufindenen Elemente wirklich gefunden worden sind.
Die Genauigkeit, im Englischen precision, zeigt hingegen wie viele der gefundenen Elemente wirklich richtig sind.
Diese werden oft gemeinsam verwendet um die Güte eines Verfahrens zu zeigen, da eines wiedergibt wie viele der gefundenen Elemente richtig waren und das andere wie viele Elemente zu finden waren.
Und zuletzte die Wortfehlerrate abgekürzt mit WER für Word Error Rate, zeigt die Anzahl der Veränderungen in einem Text geteilt durch die Anzahl der Wörter in dem Text.

Eines der ersten relativ verlässlichen Verfahren nennt sich BLEU, was für Bilingual evaluation understudy steht, und wurde 2002 im Rahmen des jährlichen Treffens der Association for Computational Linguistics vorgestellt.
Dieses Verfahren basiert auf der Annahme, dass ein maschinell übersetzter Text besser ist, je ähnlicher er zu einem professionel übersetzten Text ist.
Deshalb vergleicht es einen maschinell übersetzten Text mit professionell übersetzten Referenztexten indem es beide in N-Gramme aufspaltet und vergleicht, welche in beiden vorkommen.
Ein krudes Beispiel ist der Kandidat hier rechts oben, welcher trotz fehlender Sinnhaftigkeit eine Genauigkeit von 100Um diese Situation zu vermeiden nimmt BLEU immer die niedrigere Anzahl der Vorkommnisse der beiden Texte. The kommt im Kandidat sieben Mal vor, in Referenz 1 jedoch nur 2 Mal, weshalb die maximale Genauigkeit 2/7 sein kann.
Die Anzahl der Bausteine ist hierbei nicht unbedingt erforderlich.

Das Verfahren verwendet einige weitere Vorgänge, wie einer Minimallänge von einem N-Gramm, da sonst kurze Wörter besser bewertet werden würden, um einen Wert zwischen 0 und 1 zu ermitteln.
0 bedeutet, dass die Texte komplett verschieden sind, während 1 bedeutet, dass es die selben Texte sind.
In der Praxis ist ein Wert zwischen 0,4 und 0,6 wünschenswert.
Wie man in der Tabelle rechts unten sehen kann, ist BLEU sowohl bei Monogrammen als auch bei Tetragrammen in der Lage maschinell übersetzte von professionell übersetzten Texten zu unterscheiden.
Empirische Forschung legt jedoch nahe, dass Tetragramme die besten Ergebnisse liefern.
BLEU ist zwar sprachunabhängig, doch liefern verschiedene Verfahren in verschiedenen Sprachen unterschiedliche Ergebnisse. Aus diesem Grund gibt es kein definitiv bestes Verfahren.
Auf Basis von BLEU wurden einige weitere Verfahren entwickelt welche darauf aufbauen. NIST zum Beispiel bewertet N-Gramme zusätzlich anhand der Anzahl der Vorkommnisse im Text, etwas was BLEU ignoriert. Das neueste Verfahren nennt sich LEPOR, was für Length Penalty, Precision, n-gram Position difference Penalty, and Recall steht. Also werden Texte schlechter bewertet, wenn sie kürzer oder länger als die Referenztexte sind und wenn N-Gramme eine andere Position im Text haben, wobei Genauigkeit und Trefferquote miteinbezogen werden.
Diese Unterschiedlichen Algorithmen sind abhängig von der Sprache besser oder schlechter zur Evaluierung geeignet.